2016

Network-Based Statistical Methods for the Analysis of Stock Returns

Zidan Wu
University of Rhode Island, zidan_wu@uri.edu

NETWORK-BASED STATISTICAL METHODS FOR THE ANALYSIS OF STOCK RETURNS

BY

ZIDAN WU

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

STATISTICS

UNIVERSITY OF RHODE ISLAND

2016

MASTER OF SCIENCE THESIS

OF

ZIDAN WU

APPROVED:

Thesis Committee:

Major Professor Natallia Katenka

Gavino Puggioni

Correy Lang

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2016

ABSTRACT

To maximize returns and diversify a financial portfolio, the stock price market participants have always been interested in learning associations of stock price returns for different companies. Five primary goals of this thesis are: (1) to evaluate and infer associations of stock returns between different companies in selected industrial sectors and countries, and (2) to identify groups of companies that exhibit the most similar stock market trends, and (3) to evaluate changes in associations between companies in time period from 2009 to 2015, (4) to forecast future return movements using selected classification methods, and (5) to explore the relationship between the accuracy of classification of stock return movements and network node properties.

This thesis analyzed daily stock price data collected from publicly available sources, Yahoo Finance, for a sample of eighty-nine selected companies from four industrial sectors and three countries (China, Germany, and the US) for a time period of seven years from 2009 to 2015. Daily prices were converted into returns and then used to compute a correlation matrix and a corresponding association network. Obtained network was employed to identify clusters of companies that exhibit similar return trends and to evaluate the relationships within and between different industrial sectors. To assess changes in associations between companies during special financial events, annually dynamic networks were created. Four classification methods, namely Linear Discriminant Analysis, Quadratic Discriminant Analysis, k-Nearest Neighbors, and Logistic Regression were built to predict price movements for all selected

companies. The relationships between classification accuracy rates and network properties were evaluated graphically.

The results of the network-based analysis showed that the companies that traded in the same stock market and/or belonged to the same industrial sector had significant associations. Specifically, Chinese companies had higher inner correlations in banking and telecommunication sectors; the US and German companies had stronger associations in banking and auto manufacturing sectors. Interestingly, the associations among companies became stronger and more companies tended to be grouped together in the network during significant financial events and in the early recovery periods. The results of classification analysis revealed the superior performance of logistic regression method compared to other three classification methods, particularly for the Chinese companies. Remarkably, companies that acted as followers and belonged to medium-size clusters with eight to thirteen neighbors in the association network were easier to classify than other companies, thereby supporting the relationship between classification and network-based methods.

Keywords:

Association network, stock price return, hierarchical clustering, logistic regression, classification, forecasting.

ACKNOWLEDGMENTS

It is my great pleasure to thank the numerous people who help and support me during my two-year study in University of Rhode Island. Their constant help and support make my two-year study extremely fascinating and memorable.

Firstly, I would like to give the deepest thanks to my advisor Natallia Katenka for her inspiration, advisory and supervision throughout my entire graduate study. Her unconditional support through all aspects in the research and non-research world. She devotes countless hours of her personal time to help me finish my research and thesis; her valuable advice helps me get a wonderful intern opportunity and a job offer; and her scrupulous spirit and keeping on improving attitude deeply affects me and will continuous impresses me in my future career and life.

I also would like to express my appreciation to my committee members Professor Gavino Puggioni, Professor Corey Lang. Their considerable professional knowledge and experience supports me both through courses in the Statistics and through research. My thanks also go to Professor Xiaowei Xu.for her great help and patience as the chair of the committee.

Through the past two years of my study, I am greatly supported by Professor Liliana Gonzalez, the head of Statistics, who gives me all kinds of invaluable academic advices and guidance. She also gives me the opportunity to work as a teaching assistant during the two-year study in Statistics program.

Last but not the lease, I would like to express my appreciation to my husband, Yao Lu, for his love, patience, and support throughout the completion of this thesis. I

am also truly in debt on my parents for their love, encouragement and support. Without their support, I would never be able to pursue my dream at University of Rhode Island.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

INTRODUCTION

A stock market is also known as an equity market, which is the market issuing and trading shares of publicly held companies. A stock market is comprised of buyers and sellers of stocks, or shares, of different companies. The main function of any stock market is to allow companies to trade their stocks and raise additional financial resources for future growth and development. The stock buyer and seller could be institutes or individuals. Naturally, maximization of the selling price and minimization of the buying price are the main aims of seller and buyer. The appeal of the stock market is strong. According to the WilmerHale 2016 IPO (Initial Public Offering) report (WilmerHale 2016), 963 companies have entered the US stock market and started to sell their companies' shares from 2009 to 2015.

Besides the befit to companies and investors, the stock market serves as a primary indicator of economic stability. The stock prices of leading companies evaluate the state of development and economy of different countries and/or industries. For example, the Standard and Poor's 500 (S&P500) index that is constructed by 500 large companies selected based on their market capitalization and common shares listed for trading on the NYSE or NASDAQ. Therefore, the S&P 500 is known as one of the most internationally recognized equity indices, and it is also considered to be one of the best of representations of the US Economy. Similarly, the Shanghai Composite Index (SHCSI) shows the Chinese companies performance in Shanghai Stock Exchange. Deutsche Aktienindex (DAX) known as a German stock index is

constructed by 30 major German companies traded on Frankfurt Stock Exchange. Unlike American S&P 500 indexes, DAX index is measured by the performance of only 30 largest German companies. Due to the small sample size selection, DAX couldn't well present the vitality of the Germany economy.

The performance of three indices, S&P500, SHCSI and DAX, from 2009 to 2015 is given in Figure 1, which shows an increasing trend for all selected countries. The slopes of the German and the US indices are significantly sharper compared to the Chinese index indicating these two countries' economy are bouncing back and gradually exiting the recession. The Chinese index is not stable since there is a sharply increasing trend from second quarter of 2014 to first quarter of 2015. The German and the US indices have similar trend patterns from January of 2009 to the end of second quarter of 2014. The Chinese index and German indices have similar trends from the second quarter of 2014 to the end of 2015. Generally, the stock market is a complex index that is influenced by performance of a subset of large companies. Analysis of associations between indices can only help to find the associations between different countries, but not specific companies.

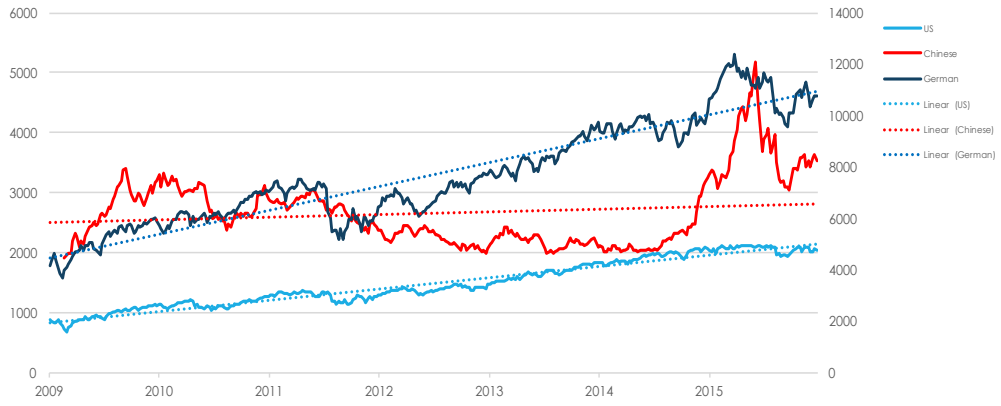


Figure 1 The stock market index performance and linear trends for three countries

Beyond discovering associations between financial indices of different countries, recovering the associations between companies based on their stock price returns is a complicated, but an important task that has received much interest among investors and financial researchers. The pair trading used in financial corporations and hedge fund is a very illustrative example of utilizing the association between different companies to gain profits.

The pair trading has been originally proposed by Gerry Bamberger and Nunzio Tartaglia's quantitative group at Morgan Stanley (Bookstaber 2007). Pair trading is done by closely monitoring two stocks whose prices are highly correlated. When the two stocks temporarily go out of sync, the trader would long the stock that is relatively lower in price and short the one that is higher in price. This way when the two stock prices converge again, the trader would benefit from his long and short positions.

The simultaneous long and short selling are widely used trading techniques in pair trading that aim at gaining profits from the relative movement of stock prices in both an upward trend market and a downward market (Ehrman 2006). For instance, consider one classic pair of companies, Coca Cola and Pepsi, two companies that produce a very similar soda product and have highly correlated stock prices. Historically, the stock prices of Coca Cola and Pepsi have had similar stock digs and highs depending on the soda market. It is expected that their price return movements to be same. However, sometimes the associations of these two companies are not synchronic. For example, the Coca Cola stock went up a significant amount while Pepsi stayed the same. Traders would buy the stocks of the long underperforming Pepsi and sell or short the stocks of the outperforming Coca. They would bet buying

Pepsi at a lower price, while selling Coca at a high price, assuming that these two companies would latter return to their historical balance. The benefit could be realized if the Pepsi stock price goes up or the Coca Cola price goes down. Wal-Mart and Target is another example of companies that sell very similar products and where pair trading could be used effectively. Thus, inferring and analyzing associations of stock returns between different companies may prove to be useful for financial researchers in both academic and corporate worlds. Additionally, it would be valuable to create reliable predictive models of price return movements and evaluate the relationships between this models and inferred associations.

Five primary goals of this thesis include: (1) evaluation and inference of associations of stock returns between different companies in selected industrial sectors and countries using correlation networks, and (2) identification of the groups of companies that exhibit the most similar stock market trends using network-based community detection techniques, (3) evaluation of dynamic changes in associations between companies in time period from 2009 to 2015 using dynamic networks, (4) forecast of future return movements using parametric and non-parametric classification methods, and (5) assessment of relationships between the accuracy of classification of stock return movements and network node properties.

To achieve these goals, daily historical stock price data was collected from publicly available national and international sources for multiple companies in selected industrial sectors in the US, China, and Germany for a period of seven years (from 2009 to 2015). Obtained prices were utilized to create a correlation matrix and characterize corresponding association network. The network properties, such as the

average node degree, network density, clustering coefficient, and average betweenness centrality, were computed for the generated correlation network. The network-based community detection method was used to find the cohesive sets of companies that exhibit the most similar stock return trends. Note that outlined network characteristics and community detection were applied to stationary network (inferred from all seven years of data) and a sequence of dynamic networks (inferred from annual data). The purpose of analysis of dynamic (annual) networks is to assess changes in associations between companies during special financial events. Since many companies depend on loans and credits, fluctuations in their stock prices may affect the stock prices of their lenders in other industries. As a result, the associations between different companies could change over time especially when facing big financial event and in the periods of early market recovery. To predict future stock return trends, parametric classification methods, including linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and logistic regression, and non-parametric k-nearest neighbor method, were applied to collected price data.

The rest of this thesis is organized as follows. Chapter 2 reviews related work in network-based and multivariate analysis of financial data. Chapter 3 outlines the data collection process and preliminary data analysis. Chapter 4 describes the network methods and predicting models used in this thesis. Results and Conclusion are summarized in Chapter 5 and Chapter 6.

CHAPTER 2

REVIEW OF LITERATURE

Application of correlation networks to the analysis of associations between different companies and/or countries based on stock market prices or exchange indices has received considerable attention among statisticians and financiers over the last ten years (Heimo et al. 2007; Nobi et al. 2014; Sienkiewicz et al. 2013; Song et al. 2011).

To analyze the association movement in stock market data, correlation graphs are often used to infer associations among the countries, industries, and/or companies, and then more sophisticated network-based techniques are employed to extract information about the global structure of the stock market. For example, correlation graph-based approach has been used to analyze the presence of both short-term and long-term association dynamic among stock exchange indices of 57 countries in the time period from 1996 to 2009 by Song (Song et al. 2011). Threshold-based correlation networks have been utilized to explore the effect of global financial crisis of 2008 on the association of stock prices in local Korean stock markets by Nobi (Nobi et al. 2014). stock returns of 96 US companies from 2005 to 2012 have been used to build a correlation network and analyze the movement of both credit and stock market during and after 2008 financial crisis by Lim (Lim et al. 2014). The maximal spanning tree of 116 US stocks returns (collected in the period from 1997 to 2000) has been constructed based on the correlation network, and then used for identification of the industry sectors of central nodes in the network by Heimo (Heimo et al. 2007).

Besides analyzing the association movement in stock market data, using historical data to create an optimal predictive model for current and future events became very popular in the new millennia (Leung, Daouk, and Chen 2000; Wang and Shang 2014; Kara, Acar Boyacioglu, and Baykan 2011; Alrasheedi 2012; Nguyen, Shirai, and Velcin 2015; Zhang et al. 2015; Peng 2015).

Specifically, to predict the movement direction of future stock price/index, many various statistical and machine learning classification models, such as Linear Discriminate Analysis (LDA), Artificial Neural Network (ANN), Support Vector Machines (SVM), have been applied to financial data. For example, the relationship between historic records of three indices, SP500, FTSE, Nikkei, and short term interest rate data in the period from 1967 to 1995 have been utilized to build models and test the predictive strength of constructed classification models via LDA, Logit and Probit Logistic Regression (Leung, Daouk, and Chen 2000). Wang and Shang converted the historic close price, high price and low price into different predictors, and applied the Lease Squares Support Vector Machine to forecast the direction of Chinese Security Index 300, as well, the LDA, QDA and Neural Network served as comparing model (Wang and Shang 2014). Kara and Boyacioglu used the the ANN and SVM model to predict the movement direction of daily Istanbul Stock Exchange National Index 100 (Kara, Acar Boyacioglu, and Baykan 2011).

Most of the projects listed above rely on historical data of only one type (e.g., prices, returns, etc.). More recently, other sources of information have been used to build models with better performance. For instance, to predict the stock price direction of one large industry corporation in Saudi Arabia, the authors have used the historical

price data, volume data, crude oil price and indices, Dow Jones and Saudi Index (Alrasheedi 2012). The sentiments from social media have been used to build a predictive model for stock price movements by Nguyen et al. (Nguyen, Shirai, and Velcin 2015). Recently, Zhang and Li have proposed a new stock movement predictive model that has shown a superior performance of energy sector. Proposed model incorporated market capitalization of multiple companies and sentiments from Twitter associated with these companies (Zhang et al. 2015). Similarly, financial data collected from Bloomberg has been used as an additional predictor for a neural network model forecasting stock price movement (Peng 2015).

To predict the future stock price return or volatility, many time series models, such as Autoregressive (AR), Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional Heteroscedasticity (GARCH), have been proposed. The time series model normally uses the previously observed values to predict the future output. For instance, the output of autoregressive model at time t is a linear regression of its own previous values. The AR(2) model uses previous two-day records, that is the value at time t-1 and the value at time t-2, as predictor-one and predictor-two to forecast the output at time t. In this study, instead of using the time series notion to build the optimal predictive model, the focus is on comparing the predictive performance of different objects by using the same predictors.

Unlike previous research that focused on exploring the changes of association of global and local exchange indices, the research in this thesis focuses on evaluating association among a subset of leading companies that comprised multiple industries (banking, communication, manufacturing, and pharmaceutical) in three countries with

high market activity (USA, Germany, and China) located on three different continents (North America, Europe and Asia). Additionally, the correlation based threshold network is used to detect and clearly represent the associations between companies and to discover the clusters of companies following the most similar trends. Besides the static network, the annual dynamic network is employed to assess changes in associations between companies in the time period from 2009 to 2015. Also, four classification models are created and compare the prediction performance of different objects across 3 different countries under the same treatment and standard. The six years' historical data in the period from 2009 to 2014 are utilized as predictors to create the classification models, and the accuracy rate is used as a justification to evaluate the performance of different models. Finally, to find the relationships between classification accuracy rates and network node properties, graphical tools are applied here to evaluate the relationship, such as the scatter chart and level plot.

CHAPTER 3

DATA DESCRIPTION AND PRELIMINARY ANALYSIS

In order to perform a robust and an accurate data analysis that would bring valid insights, it was essential to go through three data processing steps: data collection, cleaning, and formatting. Section 3.1 explains how the data was initially obtained from a large financial data source provider, Yahoo Finance. Section 3.2. Illustrates an extensive data cleaning process that was applied to improve the data quality. Specifically, Section 3.2 focuses on situations with missing data values or differences in trading dates. Additionally, this section describes how all collected stock prices were converted to the US dollars to ensure the stock returns to be comparable across different companies, industry sectors, and countries. Finally, Section 3.3 provides the results of an extensive preliminary analysis in order to understand the data and to present a reader with a general overview of the data.

3.1 Data Collection

Public daily closing stock prices recorded between Jan 2nd 2009 to Dec 31st 2015 were collected from Yahoo Finance for a subset of leading companies from multiple industrial sectors (banking, communication, manufacturing, and pharmaceutical) in three countries with high market activity (USA, Germany, and China). Closing prices were chosen as a proxy for the most accurate and commonly used measures of stock price values when financial data was collected on a daily basis.

The distribution of the companies among four industries in three countries is as follows in Table 1.

Sectors	*Country*		
	Germany	*USA*	*China*
Banking	5	10	10
Communication	3	11	3
Manufacturing	7	5	6
Pharmaceutical	7	9	13
Total	22	35	32

Table 1 Companies distribution within 3 countries and 4 industries.

Table 1 shows that the number of leading companies in communication sector in Germany and China is much smaller than the number of the US communication companies. The US telecommunication market is a liberalization market and it was subdivided into 3 main classes: one class is the large integrated telecom companies, such as AT&T, Verizon and Sprint, providing both wireless and wireline service; second class is the large wireless companies, such as T-Mobile, and the third class is the regional wireline companies, such as Frontier, CenturyLink, etcetera. However, the Chinese telecommunication system is different from the US system, it was divided by 3 companies that the Chinese government possesses. The number of leading US Auto-manufacturer is small due to the high concentration of auto market, where the head 5 manufacturers produce over 13 auto brands and occupy over 35% of the US auto-sale market.

The total of 35 US companies, 22 German companies, and 32 Chinese companies was selected as the dataset for this research. The complete list of company names, countries, and stock symbols is recorded in Appendix A.

In this study, all selected US companies are trading in the largest stock exchange: New York Stock Exchange (NYSE) and using the US Dollars as their trading currency; all selected German companies are trading in Frankfurt Stock Exchange (FSE) market in Frankfurt and they are using Euro as their trading currency; thirty selected Chinese companies are trading stocks in Chinese local stock market, Shanghai Stock Exchange (SSE), using CNY as their trading currency, and three Chinese communication companies, China Mobile Ltd, China Unicom and China Telecom Corporation, are trading their company shares in NYSE as American depositary receipt (ADR) using US Dollars as their exchange currency. It is worth noting that ADR is a stock that is traded in the US but represents a certain number of shares for a foreign stock.

Analyzing stock prices in different currencies can cause potential problems since the currency exchange rates are floating rates and can change rapidly. If the currency rates diverge greatly in a certain period of time, the analysis of the stock prices would be thus inaccurate for foreign stocks. This will compromise our ability to extract clean financial time series from the stock prices. In order to eliminate differences in currencies of obtained stock prices and to explore the associations between companies, all stock price currencies were converted to the US Dollars. The daily currency exchange rates were collected from Oanda Corporation, a Canadian foreign currency exchange company, for a period of time between 2009 and 2015. After converting all currencies to the US Dollars, the datasets for all three countries were stored in the same units, the US Dollars.

3.2 Data Preparation

The project data was collected in a period time from Jan 2nd 2009 to Dec 31st 2015; however, the date when the companies were listed on the public stock market varies. Not all selected companies started trading before 2009; some companies remained private before 2009 and started trading their company shares publicly after 2009. e.g., Tesla Motors in the US, Agriculture Bank in China, and Telefónica in Germany. As a result, there are multiple missing stock price records that could potentially complicate the analysis in this project. If a company started trading after Jan 2nd 2009, the records from Jan 2nd 2009 till the first business day before the Initial Public Offering (IPO) date would be unavailable. The records during that period are absent. Before doing further analysis, missing value imputation method was employed to the data. The numeric value of one was filled in the missing value position for the following reasons.

First reason is that since the stock returns were used for further analysis in this research, feeling zeros instead of ones in the missing records positions would result in non-identifiable values of log of zero. Clearly, such values should be avoided, for example, when computing correlations and performing classification analysis.

Second reason is that one needs to keep sufficient historical information available from the data for further data analysis. For instance, three companies did not have part of the stock price records due to IPO reason. If one completely gets rid of companies (variables) with missing values, it may cause some sectors having significantly fewer companies compared to other industrial sectors, thereby making analysis of this sector less interesting and less reliable. For example, the number of automobile

manufacturing companies in the US is already small. Tesla, a new revolutionary car manufacturer leveraging new environmental friendly electronic energy instead of gasoline, did not start trading until June 2010, but expanded its market share rapidly in the last few years. There will be significantly fewer companies in selected sectors if one removes companies like Tesla from the dataset. Computationally, it is also more convenient merging different datasets without missing values. Note that to fill in missing values after company IPOs, imputation regression method was employed.

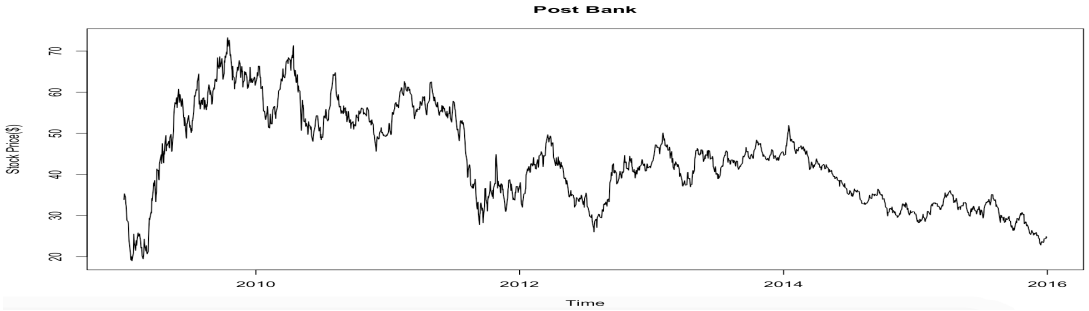
One more data issue is related to duplicated records. Stock market stops trading on holidays and different countries have their own national holidays. For example, the NYSE market is closed on President's Day and Good Friday in the US, and Shanghai Stock Exchange (SSE) market is closed for 7 days during Guoqing Festival. In the US, if the market is closed a day, it keeps no record for that day. However, Chinese market keeps the previously available record as holiday trading record in Yahoo finance. For example, during Chinese Guoqing Festival from October 1st, 2015 to October 7th, 2015, the Shanghai stock exchange market was closed, but the Yahoo finance still kept stock records from a day before holidays (September 30th 2015) and used it as real stock records for the next 7 days. Thus, the record of September 30th 2015 was used 8 times. Realistically, the records during holidays cannot be considered historical records. In order to keep more realistic data records, duplicated records in the research data were removed (e.g., 7 days from October 1st 2015 to October 7th 2015).

After removing duplicated records and solving missing values, the total of 1640 trading day was recorded for Chinese companies, 1778 days and 1761 days for German and the US companies, respectively. Combining data for all three countries

and keeping only common trading day records resulted in 1621 common trading days to be used in this research.In what follows, we present a general overview of the data to the reader. And we organized the rest of the chapter as follows: in Section 3.3.1 we describe how we use method to normalize research data and reduce the variance of the stock changes; in Section 3.3.2 we discuss the distribution of return varies for 3 countries across 7 years; in Section 3.3.3 we use Pearson correlation to analyze the create the correlation matrix and evaluate how the average relationship and volatility varies as a function of time.

3.3 Preliminary Analysis

Price chart and Candlestick Chart are commonly used plots for financial representation that are used to describe how a stock evolved over a given period of time. The following three price charts depicted in Figure 2 show how the closing prices of Post Bank, Citi Bank and ICBC changed over 7 years from the beginning of 2009 till the end of 2015.



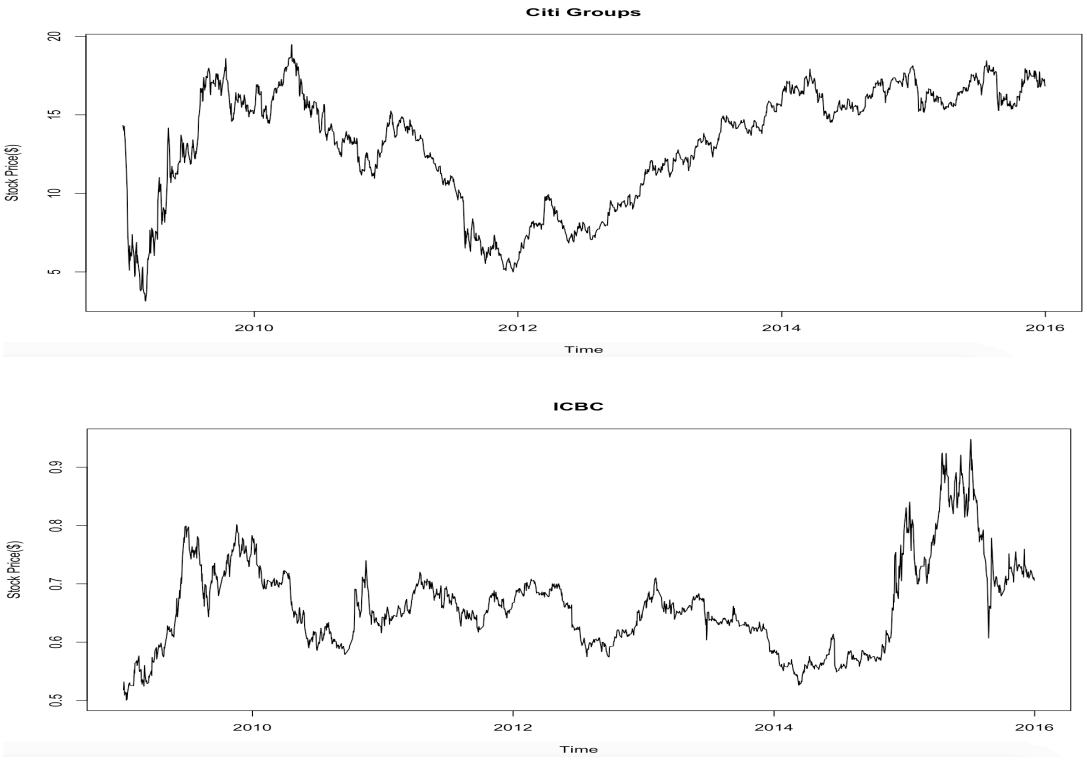


Figure 2. The price chart over 7 years for stock prices of three selected banks: Post Bank, Citi Bank and ICBC trading in Germany, the US, and China, respectively)

Figure 2 represents Post Bank having an unsatisfied continuously decreasing trend after 2010. Its stock price decreased 66% from January 2010 to December 2015. The Citi Bank had a severe depression around 2012 and a rapid recovery after 2012; and the stock prices of ICBC change in a limited range following a similar pattern of movement as Chinese stock index.

The reasons of causing stock rise or fall vary. In generally, there are three main factors that could affect stock price movements. First factor is a fundamental performance of a company, for example, the change of management, the earnings and profits related news, release of a new product, lay-off employees and etc. Second

factor is related to the industry performance and behavior of competitors. The last factor is related to the overall national economy and economic policy.

3.3.1 Returns

In this thesis, the returns were calculated and would be applied for the further analysis. Compared to stock prices, the big benefit of using stock returns instead of stock price is that the stock returns reduce the variance of stock changes, or volatility. Apparently, the stock price would have extremely large change compared to the previously recorded price especially during significant financial events. Song et al. (D.-M.Song, 2011) explored this phenomenon for several financial events including the 1997 Asian crisis, 1998 Russian Crisis and 2008 global crisis. The daily return is computed by the following equation:

$$R_i(t) = \ln S_i(t) - \ln S_i(t\text{-}1), \quad [1]$$

where $S_i(t)$ is the closing price of company i on day t.

3.3.2 Return Distribution

Eighty-nine companies were selected for this project, where each country had 20 to 30 companies. In what follows in Figure 3, the averaged returns are represented as a function of time for each country; the X-axis represent a seven-year timeline from 2009 to 2015 and the Y-axis stands for the corresponding stock returns records.

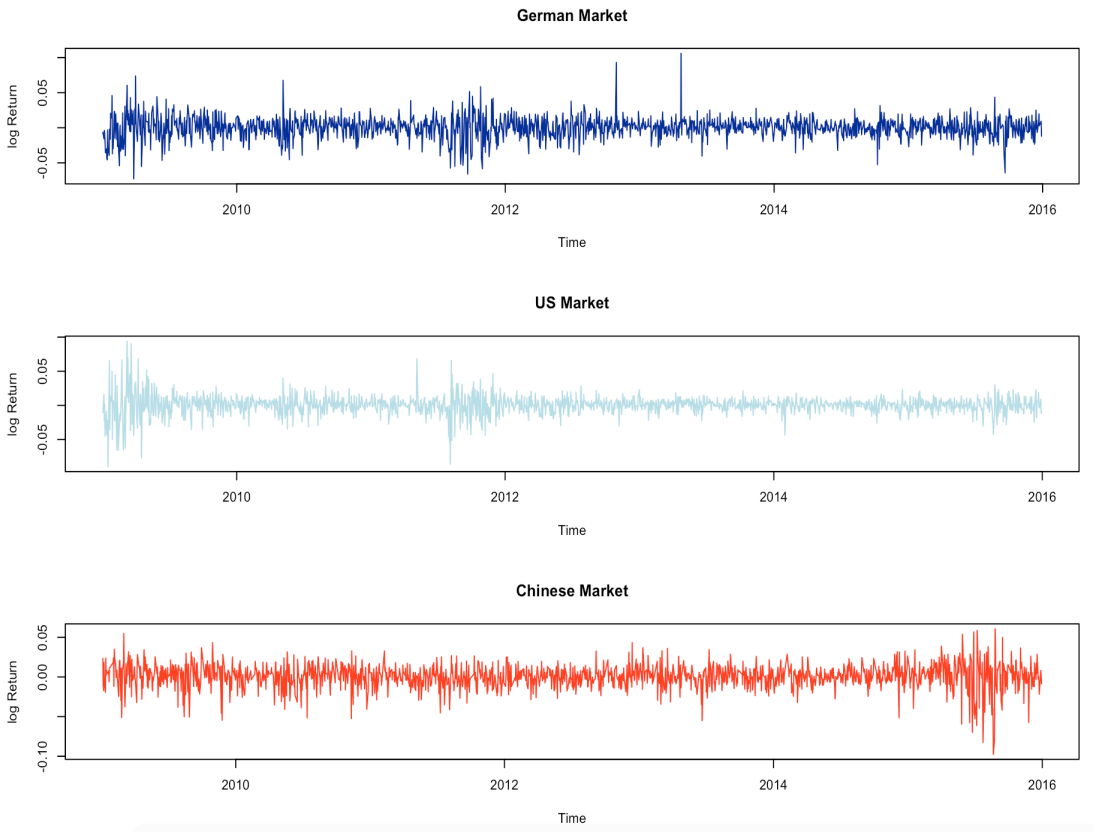


Figure 3 Average returns as a function of time. The X-axis represents 7-years timeline from 2009 to 2015; the Y-axis stands for the corresponding stock returns records.

Figure 3 shows that the three distributions of daily average returns have mean zero and. Daily stock returns change in a small range from negative 0.05 to positive 0.05. This illustrates how stock returns would normally change upward or downward in less than 5% compared to yesterday returns. Figure 3 also shows that there are three significant fluctuations around 2009, 2011 and 2015 of different intensities for different countries. This suggests that Germany, the US and China have different state effects caused by a global financial crisis in 2008. During following 2-3 years of the financial recession period, world economic market presents a declining trend, but the exact time of the recession varies from country to country. Another significant

fluctuation can be observed in 2015. In the third and fourth quarters of 2015, Chinese financial stock market, a second economy market in the world, crashed. This led to visible market fluctuations in the Western Europe stock market. There are multiple reasons causing Chinese stock market to crush including slowing down industry, the devaluation of Chinese currency, and the government froth cleaning strategy, to name a few.

		2009	***2010***	***2011***	***2012***	***2013***	***2014***	***2015***
	Germany	-0.275	-0.126	-0.519	-0.211	-0.235	-0.693	-1.073
Min	**US**	-0.724	-0.129	-0.455	-0.214	-0.699	-0.681	-0.698
	China	-0.69	-0.688	-0.917	-0.481	-0.460	-0.710	-0.609
	Germany	0.495	0.194	0.213	0.290	2.301	0.203	0.166
MAX	**US**	0.322	0.099	2.279	0.313	0.218	0.151	0.303
	China	0.129	0.182	0.101	0.104	0.113	0.173	0.136
	Germany	0.001	0.000	-0.001	0.001	0.001	0.000	0.000
MEAN	**US**	0.000	0.000	-0.001	0.000	0.001	0.000	0.000
	China	0.002	0.000	0.000	0.000	0.000	0.001	0.000
	Germany	0.036	0.023	0.030	0.025	0.039	0.023	0.025
SD	**US**	0.041	0.018	0.035	0.019	0.018	0.017	0.019
	China	0.030	0.028	0.027	0.020	0.023	0.023	0.038

Table 2 Summary of stock returns for three countries and across 7 years.

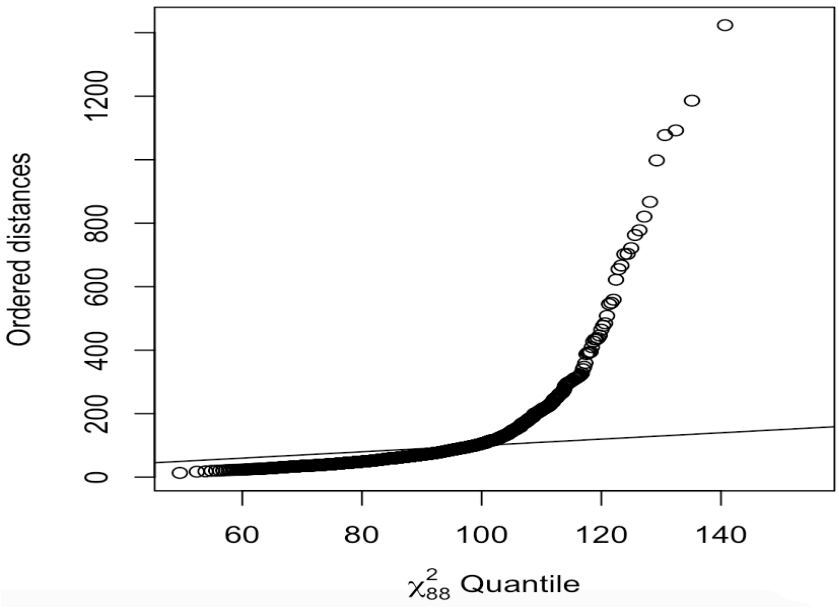


Figure 4 Chi-square plot of generalised distances for the returns

In order to preliminarily observer and visual check the normality of the returns, the Chi-square Q-Q plot is created, where X-axis is the expected distance and Y-axis is the actual distance. If returns follow a multivariate normal distribution, the points on the graph should approximately lie on a straight line, otherwise, the points are off the line. The Figure 4 represents numerous points are skewed and not on the line. Thus, one cannot infer the research data follow a normal distribution.

3.3.3 Stock Return Correlation and Annual Volatility

- **Pearson Correlation Matrix**

Pearson correlation is one of the most popular statistical measures that evaluates an association between two different numerical sets. Pearson correlation can be used to estimate the similarity between stock returns of two companies i and j as follows:

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}, \quad [2]$$

where $\{\hat{\sigma}_{ij}\}$ is a sample covariance between stock returns of company i and company j, and $\hat{\sigma}_{ii}, \hat{\sigma}_{jj}$ are the sample variances of stock returns of company i and company j, respectively. Figure 5 illustrates estimated values of the correlation coefficient $\hat{\rho}_{ij}$ between companies i and j obtained from 7 years of daily collected data. Naturally, a correlation coefficient ranges from -1 to 1 with higher positive values indicating direct linear relationship between stock returns.

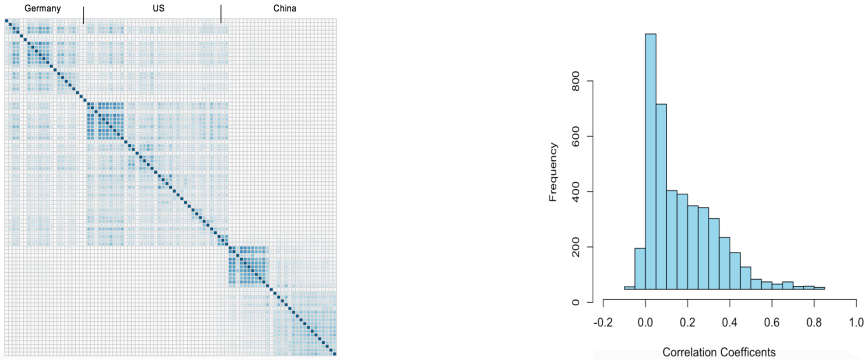


Figure 5 Correlations matrix and histogram of correlation coefficients. In the correlation matix figure (left panel), the darket color indicates stronger correlation between companies. Two figures were created using the entire 7-year period.

The summary of correlation coefficients computed from daily stock returns between different companies is illustrated in Figure 5. The left panel of Figure 5 illustrates a correlation matrix that demonstrates that the majority of correlation values are positive meaning the most companies returns followed same trends. The right panel of Figure 5 supports this conclusion that is the distribution of correlation coefficients is right-skewed with values ranging from -0.1 to 0.85.

Country	*Banking*	*Manufacturing*	*Communication*	*Pharmaceutical*
Germany	0.393	0.491	0.272	0.236
US	0.685	0.356	0.311	0.368
China	0.607	0.307	0.609	0.428

Table 3 Mean correlation across 3 countries and 4 industries.

In order to evaluate the inner correlation across 3 countries and 4 industries, we computed the average correlation coefficients and summarized the results in Table 3. One can see that the inner correlation of German Communication and Pharmaceutical is smaller than other two German sectors, especially the German Manufacturing. US banking has the largest inner correlation which equals to 0.685. As well, the Chinese communication and banking has greater inner correlation than the other two Chinese industries.

- **Returns Annual Volatility**

Volatility is also known as the variation of a stock or index, and it could estimate how risky a particular stock/ index is (Ensor,2014). Volatility could be measured using the security's standard deviation, which describes how tightly the stock prices are distributed around the mean. Because the value of standard deviation is always positive, volatility is also positive. In this section, the annual average volatility is measured as follows:

$$\sigma_{Annual} = \sigma_{SD} \quad * \sqrt{P}\ , \qquad [3]$$

where σ_{SD} is the standard deviation of a daily return, and P is the time period of return. For example, there are 232 trading days in 2009, so the related P is 232.

Left panel of Figure 6 plots the annual average correlation and volatility 89 companies under consideration. The annual average correlation shows an upward trend before 2011 reaching its peak in 2011 with value of 0.26. The rapid increase in trend can be explained due to the European sovereign debt crisis in some European countries. This financial crisis could cause an 18% increase in the average correlation. After 2011, the average correlation decreased by 50% from 2012 to 2014 and sharply increased by 61% from 2014 to 2015, which can be attributed to Chinese financial crisis.

It is well known that high volatility in stock market is related to special financial events (S. Leonidas,2012). In general, the correlation would change with volatility. In the left panel of Figure 5, one can see that the mean correlation and the average volatility have same movement trend between 2010 and 2015. During special financial events, such as European debt crisis in 2011 and Chinese stock market turbulence in 2015, the average volatility increased to 15% and 18%, respectively.

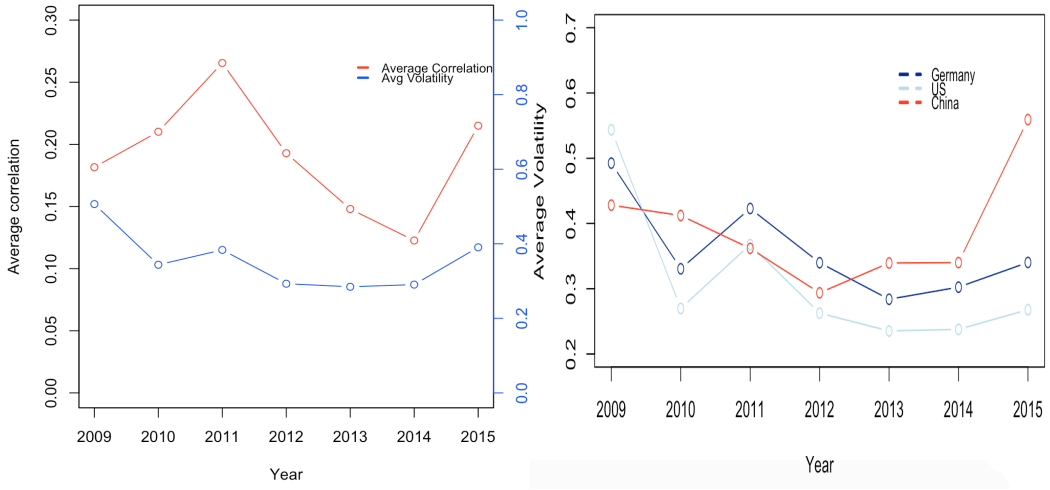


Figure 6 Left panel: annual average correlation (plotted in red) and average volatility (plotted in blue) as a function of time. Right panel: annual average volatility as a function of time for three different countries.

The right panel of Figure 6 demonstrates the annual average of historical return volatility for all selected companies in Germany, the US and China combined. This graph shows that the US and Germany volatility have same downward trend with two lines to be approximately parallel. The volatility of Germany changes to 31% in the time period from 2009 to 2015, whereas the mean volatility of US changed to 36% in the same period. The volatility of the US stock returns is greater than the volatility of German stock returns a year after the global financial crisis in 2008. In 2009, the volatility of the US returns has the largest hit compared to China and Germany. In this period the majority of the US companies relived dramatic changes. In 2009, many American companies started their recovery while others were still in recession which could explain relatively high volatility in 2009. A sharp increase in the volatility (up to 67% in one year) was also observed in 2015 in China that could be attributed to Chinese stock market turbulence. Other financial events could also contribute to significant changes in the average correlation and the average volatility, but they remained beyond the scope of this project.

The findings of the preliminary analysis present that the return has a mean zero and varies in a small range. The majority positive correlation coefficients between different companies implied that most companies' returns followed same trends and were affected by the national economy. The volatility represented the variation of returns and usually has a similar trend with the average correlation of 89 research companies. However, the volatility only refers the variation of a specific asset or index

and can not provide information on how the relationship changes within the research data. Thus, in order to fill the gap and detect the associations between companies, we will introduce and describe the network-based analysis methods and classification models in the following chapter. The main findings and conclusion of proposed study are summarized in Chapter 5 and Chapter 6.

CHAPTER 4

METHODOLOGY

The methodology used in this thesis mainly focuses on two important domains of statistical analysis. The first domain, statistical analysis of network data, is used to evaluate the relationships between selected companies based on their daily stock returns and to infer and characterize association networks. The second domain, classification analysis, is aimed at forecasting the direction of future returns of stocks using classification techniques such as linear discriminant analysis (LDA), quadratic discriminant analysis, logistic regression and a non-parametric model, k-nearest neighbors (KNN).

Specifically, Section 4.1 introduces how the association networks are built utilizing correlations computed between time series of stock return of different companies, and illustrate the principles of FDR test which aims at avoiding generating exaggerated type I error. Section 4.2 focuses on the threshold network model which is adapted to explain large and complex networks. In addition, Section 4.3 and Section 4.4 describe methods for computing network characteristics and the process of testing the significance of the computed characteristics. Section 4.5 further clarifies how the annual networks were inferred in this research. Finally, Section 4.5 focuses on the four classification methods that are used for forecasting the direction of future stock returns.

4.1 Correlation-Based Network

Statistical analysis of network data combines an area of mathematical graph theory and statistical data analysis. In proposed study, correlation-based networks are used to represent the association between independent objects.

There are two main components that comprise any network, namely vertices and edges. Formally, a network graph, or simply, graph G= (V, E) is used to represent a set of elements (vertices) and a set of interconnections (edges) between the elements, where V denotes a set of vertices (also commonly known as nodes) and E denotes a set of edges (also commonly called links). It what follows, the network graph under consideration is assumed to be a simple, undirected graph, that is a graph with no multi-edges between any elements, no edges that connect an element to itself (self-loops), and where the direction of edges is not important.

For the purposes of the proposed study, vertices (V) are defined as the leading companies selected from multiple industrial sectors (banking, communication, manufacturing, and pharmaceutical) in three countries with high market activity (USA, Germany, and China), while the edges (E) are defined as the connection between any of the vertices (companies). The correlation network of stock returns can be constructed in two steps:

(1) Evaluate the similarities between different vertices using Pearson correlations introduced and computed in the previous chapter of this thesis; and

(2) Utilize an appropriate test statistic to verify if the similarities between daily stock returns of different companies are statistical different from zero.

In Section 3.3, Pearson correlation was introduced as a measure that can evaluate the similarities in prices of different companies within the dataset. Here, ρ_{ij} is used to denote the correlation coefficient between time series of stock returns for a pair of companies i and j.

The following hypothesis tests are applied to verify the existence of significant linear relationships between possible pairs companies:

$$Ho: \rho_{ij} = 0 \quad versus \quad \text{Ha: } \rho_{ij} \neq 0, \ for\ all\ \{i,j\} \in V^{(2)}. \quad [4]$$

If a null hypothesis is rejected at a significance level of 0.05, an edge between two vertices i and j is assigned and is added to a set of edges E in the correlation network G; otherwise, no edge is assigned between these two vertices. At the end, the set of edges is comprised from edges $E = \{\{i,j\} \in V^{(2)}: \rho ij \neq 0\}\}$. The value of the corresponding test statistics is:

$$Z_{ij} = tanh^{-1}(\hat{\rho}_{ij}) = \frac{1}{2}\log\left[\frac{(1+\hat{\rho}_{ij})}{(1-\hat{\rho}_{ij})}\right]. \quad [5]$$

Assuming that a pair of stock returns follows a bivariate Gaussian distribution, under Ho the distribution of Z_{ij} is well approximated by a Gaussian random variable with mean zero and variance 1/(n-3), where n is a sample size of commonly available observations for each pair of companies (maybe different for different pairs). It is worth noting that the approximation of Z_{ij} under Ho remains valid even when a pair of stock returns departures from Gaussianity, but sample size n is large (see Hotelling 1953).

However, if multiple hypothesis tests are performed on the same data, the results from one test may be related to the results of from another test, and therefore the type I error rate may be much higher than pre-defined significance level α (the problem of

multiple testing). To address this problem, the Benjamini- Hochberg adjustment is applied here, where the p-value is adjusted based on the control of false discovery rate (FDR). The procedure of FDR is organized as follows. After finding the statistical p-values of multiple tests described by Equation 4, one needs to sort these p-values from smallest to largest, generating a sequence of $p(1) \leqslant p(2) \leqslant \cdots \leqslant p(N)$, and then declare potential edges in the network for pairs of nodes for which $p(k) \leqslant (k/N)\gamma$, where k is the k^{th} smallest p-value among the N tests. This formula means all hypothesis tests with smaller p-values than p(k) will be rejected (the null hypothesis will be rejected at the level γ) and corresponding edges are going to be assigned. Here, the level γ is a user specified value.

After the p-values are adjusted, the adjusted p-value will be compared with the significance level, here, the standard 0.05 significance level is utilized.

4.1.1 Threshold Correlation Network

Threshold correlation network approach is similar to the correlation network approach described above. It is, in fact, a simpler approach that can be very helpful if the inferred correlation network is too complex and a large percentage of edges are significantly different from zero. The threshold network ignores associations between two vertices with the corresponding correlations smaller than a pre-set threshold and preserves the associations with the corresponding correlations greater than the threshold.

Formally, we create network graph G= (V, E) following the formula [6] to determine edge set E (see A. Nobi, et al., 2014):

$$E_{ij} = \begin{cases} e_{ij}, & if\ \rho_{ij} \geq \theta \\ 0, & otherwise. \end{cases}, \qquad [6]$$

where θ stands for the correlation threshold, and ρ_{ij} represents an estimated correlation between company i and company j. The character e_{ij} denotes the edge between node i and j.

Note that the threshold network is only able to create an undirected network graph and depict if the correlations between two vertices are larger than the threshold value θ. The threshold network graph cannot describe the direction of edges. We also should be cautious about the value of threshold θ. The network will be fully connected if θ is set too small, and the network will be empty if θ is far too large. Therefore, choosing a suitable θ is one of the key elements of this part of analysis.

4.1.2 Random Graph Models

Random graph models are frequently used to test the 'significance' of characteristics in a constructed network graph. Here, we adapt two random graph methods: one is a classical random graph model originally proposed by Erdős and Rényi; and one is a generalized random graph.

The core concept of the classical random graph theory is adding successive random edges to a set of N isolated vertices. To test significance of structural characteristics of the observed network, a sequence of classical random graphs is created where each graph has the same node number and edge numbers as the observed network graph. Formally, for network graph G, 1,000 random graphs are simulated with $N_v = |\mathrm{V}|$ and $N_e = |\mathrm{E}|$, where N_v and N_e denote the vertex number and the edges number of the observed network graph, respectively.

Similarly, a sequence of generalized random graphs is created where each graph has the same number of nodes and the degree sequence of as the observed network graph (see Section 4.1 for more details).

After obtaining the simulated graphs, one can calculate the distribution of structural network characteristics computed for each random graph. The examples of structural network characteristics include but not limited to a graph density, an average betweenness centrality, an average vertex degree, and a clustering coefficient. All these characteristics are introduced in the next section. The distribution of a given network characteristic constructed from simulated classical or generalized random graphs can be considered a reference distribution that one can use to examine how likely the characteristic of the observed graph is under this distribution. This likelihood can be used then to assess the significance of the observed network characteristic compared to a random graph structure.

4.1.3 Network Characteristics

To detect and describe the structure in an observed network graph, the following network characteristics are computed: vertex degree, graph density, clustering coefficient, and betweenness centrality.

- **Vertex Degree distribution**

Vertex degree is defined as a measure of vertex connectivity in a given network. In the network graph G=(V, E), the degree of a vertex v, denoted as d_v, is the number of edges in graph G incident to the vertex v. Hence, the degree of a company is the number of other companies associated with this company, or the number of companies the price returns of which have a strong linear relationship with

the price returns of a given company. Often, the degree distribution is used as a fundamental property of network graph, because it could be easily computed and interpreted as the representation of the network connectivity.

- **Graph Density**

To evaluate whether or not a given vertex is the 'central' vertex in the network graph, the network density characteristic is included in the proposed analysis.

The global density characteristic of a graph could be defined as the frequency of realized edges relative to its potential edges, and it lies between zero and one (E. D. Kolaczyk, 2014). The number of potential edges in an undirectional graph G= (V, E) with no self-loops and multiple edges is equal to $N_v*(N_v-1)/2$. Thus, the density of graph G can be defined as:

$$den\ (G) = \frac{|E|}{\frac{|V|(|V|-1)}{2}}, \qquad [7]$$

where E is the number of realized edges in G and V is number of vertex.

Similarly, the density of a sub-graph H= (V_H, E_H) is:

$$den(H) = \frac{|E_H|}{|V_H|(|V_H|-1)/2}, \qquad [8]$$

which can be used to measure how close is the subgraph H to a clique.

- **Clustering coefficient**

Clustering coefficient (cl) is a network characteristic, similar to a graph density. Both a clustering coefficient and a graph density are used to describe the cohesive properties of a given graph.

Clustering coefficient (cl) is a measure of the frequency with which connected triples 'close' to form full triangles in the undirected graph G= (V, E). It could be computed using equation [9]:

$$cl_T(G) = \frac{3\tau_\Delta(G)}{\tau_3(G)}. \qquad [9]$$

where $\tau_\Delta(G)$ is one third of the number of triangles in graph G, and $\tau_3(G)$ is the number of connected triples.

The clustering coefficient also can be computed locally. The local clustering coefficient of a node may help to determine whether neighbors of a node also connected and how close they are to forming a clique. The clustering of a vertex i in graph G could be obtained using the following equation:

$$\mathrm{cl}_T(i) = \frac{2|\{e_{jk}: v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i\text{-}1)}, \qquad [10]$$

where the v_j and v_k are the neighbors of vertex i, and e_{jk} represents the edge between node j and k. The k_i denotes the number of neighbors of node i.

The value of cl_T is also called transitivity of the graph and widely used in the social network literature. In the social network, the clustering can indicate how likely one person's friends befriend each other. Similarly, the local clustering coefficient in this thesis suggests how much the associated companies of a specific node are also highly correlated with each other. The global version of the cluster coefficient evaluates the overall level of clustering in the network. It presents how likely corporations in the data are interconnected and how likely they form communities.

- **Betweenness centrality**

To investigate and quantify the 'importance' of vertices in a network graph, a betweenness centrality measure is used. There are several variants available, but the most commonly used definition of betweenness centrality is:

$$c_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}, \qquad [11]$$

where $\sigma(s,t|v)$ is the total number of shortest paths between s and t that pass through v, and $\sigma(s,t)$ is the total number of shortest paths between s and t (include both pass or not pass the vertex v). If the vertex v has the largest $c_B(v)$ which means this vertex has large probability being a central vertex in the network graph. Here, the average of all vertices betweenness centrality coefficients were computed at different threshold.

4.2 Network Community Detection

In order to identify groups of companies that exhibit the most similar stock market trends, agglomerative hierarchical clustering method is utilized in this thesis. The reason of using the agglomerative hierarchical clustering instead of the divisive hierarchical clustering is that the former strategy detects and aggerates the similar companies together until only one cluster left; the latter one merges all companies together, and gradually detaches the dissimilar corporations. This thesis is more interested in the similarities between different companies; thus, the agglomerative hierarchical clustering method is employed.

The agglomerative hierarchical clustering algorithm works as follows. In the beginning, the algorithm places every element into its own cluster. Next, according to the hierarchical clustering principle, two clusters with the most similar properties

merge thereby creating a new cluster. On each step of the algorithm, two of the most similar clusters merge. The process continues until all clusters are merged and all elements belong to only one cluster.

In this thesis, the roles of elements play the companies in selected sectors and countries. Many methods can be utilized to measure similarities between companies. Here, Euclidean distances are utilized to measure the similarities between stock price returns.

Formally, Euclidean distance is defined as the length of a straight-line distance from one point to another point in Euclidean space. Suppose p is one company (point) with n return records, then p= ($p_1, p_2, \ldots, p_n$), and q is another company (point) with n return records written as q= ($q_1, q_2, \ldots, q_n$) in Euclidean n-space. The distance between two companies (points), p and q, can be computed using the formula [12]:

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \mathrm{d}(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1\text{-}p_1)^2 + (q_2\text{-}p_2)^2 + \cdots + (q_n\text{-}p_n)^2}. \qquad [12]$$

Note that d(**p, q**), the distance between two points, p and q, is an undirected line segment connecting these two points. In this project, the data contains 1621 observations (n=1621) for each of 89 companies, so the pairwise distances between each pairs of 89 vectors of size n need to be computed.

Given the matrix of computed Euclidean distances, one can apply the hieratical clustering algorithm. First, all companies are initially separated in their own clusters. Next, two companies with the minimal distance are merged. As this new cluster is formed, the distance between two clusters with multiple elements needs to be computed. In this thesis, the complete linkage is used to calculate such distance [13]:

$$d_{AB} = \max_{p \in A, q \in B} (d_{\boldsymbol{pq}}), \qquad [13]$$

where d_{AB} is the distance between two clusters A and B, and d_{pq} is the distance between each vectors of stock returns p and q in clusters A and B. This way, one can use the maximum distance of d_{pq} as the distance between two clusters A and B. After computing the distances between new clusters, clusters with the $\min d_{AB}$ are merged. As mentioned earlier, this process continues until all companies are merged in one cluster. As a result, a dendrogram tree is constructed that illustrates the arrangements of merged clusters. Usually, clusters are defined by cutting branches off the dendrogram tree at a specific value of heights that is a closeness measure of different companies or clusters. Note that cutting the dendrogram tree at different heights will result in different clustering solutions.

There are other options to compute the distance between two clusters including single linkage, average linkage, to name a few. The reason of utilizing the complete linkage instead of using, for example, the commonly used single linkage is that the research data is highly correlated (see the preliminary results in Chapter 2). The single linkage clustering has a disadvantage on analyzing high correlated data. Following the single linkage algorithm for highly correlated data will result in the situation where on each new iteration the existing cluster merges one new observation with the closest similarity with the created clusters. At the end, if one cuts the tree at a specific value, the tree will be divided into two main parts, one is a cluster, and another one is a set of separated self-clustered companies. The result does not have much of explanation value. Yet, the complete linkage clustering does not have this limitation and for the highly-correlated data this type of linkage forms some small cliques first, and then uses the dissimilar companies in each clique to compute the similarities between two

clusters. If one cuts the tree at a specific value, the tree will be divided into different clusters containing companies that have inner similarities. Thus, the complete linkage clustering is a more suitable technique to achieve one of the research goals of detecting the companies with similar stock return trends.

4.3 Dynamitic Networks

Dynamic network is adapted here for the purpose of discovering the changes in association of eighty-nine selected companies from four industrial sectors and three countries (China, Germany, and the US) over seven years from 2009 to 2015. The dynamic network is a statistical network analysis method used to describe the complex dynamic system. Compared to previously described (static) networks, for construction of which daily observations from all seven years are used, the dynamic network conceptually splits the data into a different time windows and uses data sequentially from each window to create a set of new networks.

In this thesis, dynamic analysis of network data is applied annually for, on average, 232 daily records, to explore the structural changes in association networks of companies, industries, and countries. Specifically, the data was separated by year (for example, the first record was computed using the stock return records in the period from Jan 2009 to Dec 2009), and for each year, one static correlation network is built and characterized. Note that the participants in the network can vary by years. For example, there were fewer members in the first four years compared to the rest three years, because some corporations did not start selling their stocks until a certain year, e.g., German Telef Telecommunication (2012), American Tesla Motor (2010), and Chinese Great Wall Motor (2011), etc.

In order to describe the Network graph, three network characteristics are computed across the 7 research years including graph density, average betweenness centrality, and clustering coefficient.

To omit the overlapping edges and represent the relationships between connected companies more clearly, the topology technique, the spanning tree, is applied in this section. The spanning tree is a subset of graph G, which has all the vertices covered with minimum number of edges. For example, if three vertices are inner connected and form a triangle in a graph G. The spanning tree H is a subset of the graph G, which simply connect all three vertices with two edges. Hence, there are no circle/loop in the spanning tree graph.

4.4 Classification Methods

In what follows next, the focus is on forecasting of the movement directions of stock price returns, and understanding of the relationships between a model predictive power and the network/data properties. The following four classification methods are explored including LDA, QDA, KNN, and Logistic Regression. The performance of these methods is compared based on the prediction accuracy of the directions of stock returns.

A typical approach to assessing model performance is separating the data into two sets, training dataset and test dataset. The training data is usually used to build a model and estimate the related parameters; the test data is normally used to test the performance of the developed model. Here, the dependent variable (Y) represents movement directions of stock price returns (Upward, Y=1/Downward, Y=0); the predictors are the returns of the eighty-nine research companies. The predictor X_{t-1} is

utilized to predict stock movement direction of a target company j at time t, $Y_{j,t}$. The predictor is matrix with 1620 rows and 89 columns, which also could be formally written as X_{t-1}= ($X_{1,t-1}$, $X_{2,t-1}$, $X_{3,t-1}$, $X_{4,t-1}$,…, $X_{p,t-1}$). The structure of predictors displays as following:

Obs	Return Movement of Y	Commerzbank	Postbank	Deutsche Bank	Aareal Bank	IKB Bank	…	Zhejiang Hisun	Joincare Pharm
1	$Y_{j,1}$	$X_{1,0}$	$X_{2,0}$	$X_{3,0}$	$X_{4,0}$	$X_{5,0}$	…	$X_{88,0}$	$X_{89,0}$
2	$Y_{j,2}$	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{4,1}$	$X_{5,1}$	…	$X_{88,1}$	$X_{89,1}$
3	$Y_{j,3}$	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{4,2}$	$X_{5,2}$	…	$X_{88,2}$	$X_{89,2}$
4	$Y_{j,4}$	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	$X_{4,3}$	$X_{5,3}$	…	$X_{88,3}$	$X_{89,3}$
…	…	…	…	…	…	…	…	…	…
n	$Y_{j,n}$	$X_{1,n-1}$	$X_{2,n-1}$	$X_{3,n-1}$	$X_{4,n-1}$	$X_{5,n-1}$	…	$X_{88,n-1}$	$X_{89,n-1}$

First six-year (2009-2014) stock returns (from row 1 to row 1388) are utilized to estimate the parameters in the classification model as a training set. And the last year (2015) research data (from row 1389 to 1620) are used to evaluate the performance of the constructed classification model

4.4.1 Linear Discriminant Analysis

Linear Discriminant Analysis is based on the assumption that the predictors X= (X_1, X_2, X_3, X_4,..., X_p) is drawn from a multivariate Gaussian distribution, where X has a class-specific mean and a common covariance matrix, and Y is a categorical class variable (James et al. 2013). In this project, Y is a two-level class variable that represents the movement direction; that is Y is equal to one if the return is positive ('Upward' movement), and Y is equal to zero ('Downward' movement) otherwise. The conditional distribution of X given Y=1 and given Y=0 is denoted as f_1(x)=f(X=x|Y=1) and f_0(x)=f(X=x|Y=0), respectively, and can be computes as follows:

$$f_1(X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \quad ,$$

$$f_0(X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X-\mu_0)^T\Sigma^{-1}(X-\mu_0)\right) \quad , \qquad [14]$$

where p is the dimension of random variable X, Σ is a covariance matrix, common for both classes, and μ_1, μ_0 are means of X for class 1 and class 0, respectively.

The optimal Bayes classification rule could be written as follows:

$$\hat{Y}\left(d_x(X)\right) = \begin{cases} 1, d_x(X) \geq \log\left(\frac{\pi_0}{\pi_1}\right) \\ 0, d_x(X) < \log\left(\frac{\pi_0}{\pi_1}\right) \end{cases}, \qquad [15]$$

where π_0 and π_1 are prior class probabilities, and $d_x(X)$ is a linear function which projects p-dimensions vector X onto one dimension with maximum class separability:

$$d_x(X) = \frac{\log\left(f_1(X)\right)}{\log\left(f_0(X)\right)} = (\mu_1-\mu_0)'\,\Sigma^{-1}\left(X-\frac{1}{2}(\mu_1+\mu_0)\right). \qquad [16]$$

The unknown parameter π_0, π_1, μ_1, and μ_0 are estimated from training set:

$$\widehat{\pi_0} = N_0/N, \qquad \widehat{\pi_1} = N_1/N,$$

$$\widehat{\mu_0} = \sum\frac{X_0}{N}, \qquad \widehat{\mu_1} = \sum\frac{X_1}{N},$$

$$\hat{\Sigma} = \frac{1}{N-2}\sum_{c\in\{0,1\}}\sum_{Y_i\in c}(X_i-\widehat{\mu_c})(X_i-\widehat{\mu_c})', \quad [17]$$

where N is the total number in the training set and N_1 and N_0 are the numbers in class 1 and 0, respectively. Character c denotes the different class of Y, here, Y has two-class.

4.4.2 Quadratic discriminant analysis

Quadratic discriminant analysis (QDA) is similar to LDA in basic concepts, predictors X= (X1, X2, X3, X4,…, Xp) follows a multivariate Gaussian distribution. However, the QDA does not assume the covariance is homogeneous, it allows

different two groups within the data having different covariance, Σ_1 and Σ_0. The conditional distribution of X given Y=1 and given Y=0 is denoted as f_1(x)=f(X=x|Y=1) and f_0(x)=f(X=x|Y=0), respectively, and can be computes as follows:

$$f_1(X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_1|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)\right)$$

$$f_0(X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_0|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X-\mu_0)^T\Sigma_0^{-1}(X-\mu_0)\right), \quad [14]$$

in addition, in QDA the decision boundary is quadratic.

The QDA function and the classification rule could be written as:

$$d_1(X) = -\frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(X-\mu_1)'\Sigma_1^{-1}(X-\mu_1) + \log\pi_1,$$

$$d_0(X) = -\frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(X-\mu_0)'\Sigma_0^{-1}(X-\mu_0) + \log\pi_0, \quad [18]$$

And

$$\hat{Y} = \begin{cases} 1, d_1(X) \geq d_0(X) \\ 0, d_1(X) < d_0(X) \end{cases}, \quad [19]$$

where the unknown parameter μ_1, and μ_0 are estimated from training set using the formula [17], and unknown covariance for each class could be estimated using

$$\widehat{\Sigma_1} = \frac{1}{N_1-1}\sum(X_i-\widehat{\mu_1})(X_i-\widehat{\mu_1})'$$

$$\widehat{\Sigma_0} = \frac{1}{N_0-1}\sum_{i\in 0}(X_i-\widehat{\mu_0})(X_i-\widehat{\mu_0})'.$$

4.4.3 K Nearest Neighbors

K-Nearest Neighbors (KNN) classification is a distance based, non-parametric classification method that does not require feature vector X to follow any particular distribution. The steps of employing a KNN algorithm are:

(1) label all observations in the training set and compute the distances from each observation in the testing set to each observation in the training set, in a pair-wise fashion;

(2) sort the distances and detect k of the closest neighbors for each point in the test set;

(3) use a majority vote to label the points in the test set using the labels of k closest neighbors in the training set.

The Euclidean algorithm is utilized here to calculate the distances between all objects in the test set and all other objects in the training set. The procedure of computing Euclidean distances in KNN has some differences from building a distance matrix in Hierarchical clustering. Like for any classification algorithm, for KNN, data is divided into two sets, a training set and a test set. Let p to be a vector of price returns in the test set, p= (p_1, p_2,…, p_p), and q to be a vector of price returns in training set and q= (q_1, q_2,…, q_p). Every vector in the training set are labeled. In order to correctly label vectors in the test set, the Euclidean distance are computed using the formula [12]:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1\text{-}p_1)^2 + (q_2\text{-}p_2)^2 + \cdots + (q_\mathrm{p}\text{-}p_\mathrm{p})^2}, \qquad [12]$$

where the distance between two vectors p and q is an undirected line segment connecting these two points. Here, the training dataset has 1388 vectors and each

vectors has 89 numbers, at the same time, the testing data has 232 vectors. The Euclidean space has 89 dimensions.

The nearest vectors X_i would be obtained by sorting the computed distance values. And if there are no ties, the classification (label) for each row of the test set is voted by the majority neighbors in the K candidates, and if there are ties for the K nearest vectors, all candidates will be included to vote. For example, if more than one-half of k neighbors around vector p have label 'Upward' movement ($\hat{Y} = 1$), the p will be assigned the label 'Upward', otherwise, the p will be assigned the label 'Downward' movement ($\hat{Y} = 0$).

4.4.4 Logistic Regression

Logistic regression model is another classification method. It can be considered a special case of the generalized linear regression model with a binary dependent class variable Y that is used for prediction of the probability of Y to belong to a particular class. In this thesis, variable Y represents the movement direction of a price return; Y equals to 1 ('Upward' movement) if the corresponding daily return is positive, and Y equals to 0 ('Downward' movement) otherwise.

Formally, the logistic regression model can be written as follows:

$$p(X) = \frac{e^{X\beta}}{1+e^{X\beta}}, \qquad [20]$$

In Equation [20], p(X) is an estimate of how likely the dependent variable Y is equal to one ('Upward' movement); X is the predictor matrix, and β is the vector of corresponding coefficients. If data have N trading days and p predictors, X will be a N*(p+1) matrix, where $X = (1, X_1, X_2, X_3, X_4, \dots, X_p)$, and β is a (p+1) vector of coefficients. The probability value, p(X) , is positive and in the range between 0 and 1.

Generally, a threshold of 0.5 is applied to determine to which class an object belongs, that is and if p(X), Y will be equal to one and the 'Upward' movement label will be assigned, otherwise, Y will be equal to zero and the 'Downward' movement label will be assigned.

The vector of coefficients, β, in the logistic function is estimated by maximizing the likelihood function:

$$\ell = \prod_{i=1}^{n} \prod_{c=0}^{1} p(X_i)^{Y_c} (1\text{-}p(X_i))^{1-Y_c} \quad , [21]$$

where X_i is the i^{th} observation in the training data and Y_c denotes the correspond observed Y class(Y=1 or Y=0). The probability of being 'Upward' is $p(x_i)$, and the probability of being 'Downward' is $1\text{-}p(x_i)$.

4.5 Evaluation of Model Performance

Since multiple models have been applied for the prediction of price returns movements, their individual performance is of interest in this study. To evaluate the performance and validate the results of prediction, the confusion matrices and accuracy rates are computed.

Confusion matrix is a table that is broadly used to visualize the performance of any given classification algorithm. The columns in the confusion matrix represent the predicted class while each row in the table displays one of the true classes. The confusion table is constructed as follows:

		Predicted Condition	
		True	False
True Condition	True	true positive (TP)	false negative (FN)
	False	false positive (FP)	true negative (Sienkiewicz et al.)

Table 4 The confusion table layout

The accuracy rate can be obtained using the following equation:

$$\text{Accuracy rate (ACC)} = \frac{TP+TN}{P+N}, \qquad [22]$$

where TP denotes the number of positive responses correctly labeled (true positives), and TN represents the number of negative responses correctly labeled (true negatives), as well, the sum of all positively labeled responses (P) and all negatively labeled responses (N) is the given data set. It is worth noting the apparent accuracy can be calculated by using Equation 22 and the training dataset. And the true accuracy is computed by using the test dataset.

In this chapter, selected methods from two areas of statistics, Statistical Analysis of Network Data and Classification Analysis, were introduced. The numerical results of analysis and the conclusions will be presented in the following Chapter 6 and Chapter 7.

CHAPTER 5

RESEARCH RESULTS

Given the statistical methods introduced in the previous chapter, this chapter utilizes them to investigate the structure of the correlation networks inferred from price returns data (previously described in Chapter 2) and the predictive power of several classification models for corresponding price return movements. At the end of the chapter, the relationship between the accuracy of classification and the network node properties is explored.

Specifically, this chapter is organized as follows. Section 5.1 presents the correlation network inference and the analysis of its structure. The following network properties are computed the node degree distribution, density, clustering coefficient, and betweenness centrality. Section 5.2 describes the clustering analysis and identification of groups of companies that exhibit the most similar stock market price return trends. Furthermore, Section 5.3 evaluates the changes in associations between companies annually in the post-crisis time period from 2009 to 2015. Section 5.4 outlines and evaluates the predictive power of four selected classification methods. Finally, Section 5.5 addresses the relationship between the accuracy of classification on stock return movements and network node properties.

5.1 Correlation-Based Network

To detect the hidden associations between different companies in selected dataset, a correlation network G is inferred from stock market price data (see Chapter 2 for

data description). In the graph G, nodes represent companies and edges represent sufficient correlations between vectors of stock market price returns computed over period of seven years. In total, there are 89 companies represented as nodes and 3961 possible edges in the network. The existence of a significant association between these companies has been verified by testing a set of the following hypothesis:

$$Ho: \rho_{ij} = 0 \quad versus \quad \text{Ha: } \rho_{ij} \neq 0, \ for\ all\ \{i,j\} \in V^{(2)}. \ [4]$$

The results of the tests combined with the FDR analysis, show that 2809 correlations are statistically different from zero at the 0.05 level of significance. This implies that the number of edges in the constructed network is 2809 out of 3961, which is far too large for both visualization and interpretation.

For illustration, the left panel of Figure 7 presents the network with all potential 3961 edges, and the right panel of Figure 7 depicts the network with the 2809 significant edges only. One can see that the graph illustrated on the left panel is denser than the graph illustrated on the right panel, even though the network densities of the two graphs are 1 and 0.709, respectively. However, the associations between different vertices are not clear. It is still difficult to infer any meaningful conclusion by reading over complex network graph with too many overlapped edges. One research goal of this thesis is using network analysis method to evaluate the associations between different companies and industries. The overlapped edges clearly cannot contribute much to achieving this research goal.

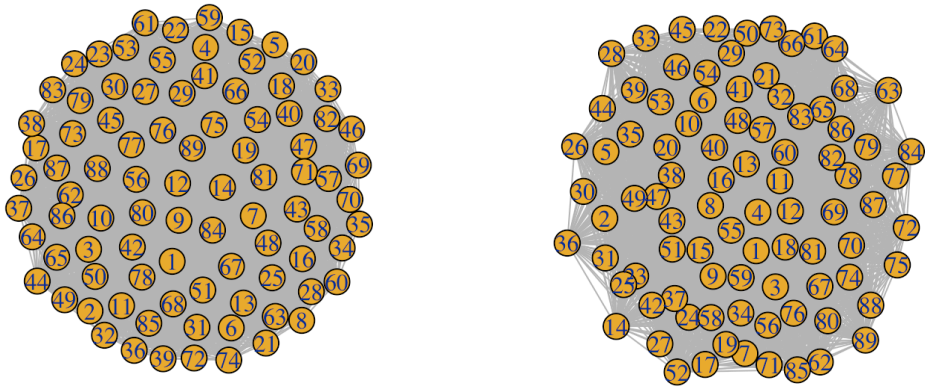


Figure 7 Left panel: Network graph with all potential edges. Right panel: Correlation network has significant edges only.

5.1.1 Threshold Network

As explained before, a substantial number of associations between pairs of corporations is significant at 5% level. It appears that the inferred network graph is extremely dense if one depicts all significant edges in the graph. Alternatively, to clarify the most important associations between different companies, the threshold network, described in Chapter 3, can be utilized.

The choice of the threshold value is of particular importance in this analysis. It is worth noting that if the threshold is set too high, only a few extremely influential associations can be present; while a large number of 'sub strong correlated' edges will be absent; alternatively, if one set the threshold very low, too many of 'less important' edges will present in the network. Here, in order to determine the suitable threshold, three important network characteristics including graph density, clustering coefficient, and betweenness centrality will be utilized.

- **Threshold Value Selection**

The left graph in Figure 8 demonstrates that the graph density decreases as the threshold value for correlation increases. The weak correlations are cut off as the threshold changes from 0 to 0.4; at the same time, the graph density descends from 0.962 to 0.097. If the graph density is very close to zero, it implies that the network graph is an empty graph with the number of edges close to 0.

The average betweenness centrality and the overall clustering coefficient are shown in the right panel of Figure 8. One can see that the clustering coefficient has a decreasing trend until it encountered the first drop at threshold θ=0.06. Then the clustering coefficient begins to increase and then climbs the first peak at θ=0.179. It is not difficult to conclude that the edges with insufficient correlation are very unlikely to form triangles and will be cut off first. At the same time, edges with stronger correlation values tend to form triangles with their connected neighbors (causing clustering coefficient to increase) will remain in the network. If threshold θ is set to even larger values, the inner-cluster edges will be removed (causing cl to decrease). When the threshold θ equals to 0.415, the network graph has a low clustering coefficient but high average betweenness coefficient, this results illustrates that few vertices in the graph have extremely high centrality compared with other vertices. When the threshold equals to 0.415, the corresponded graph density is 0.085, meaning that 8.5% of potential undirected edges will actually present.

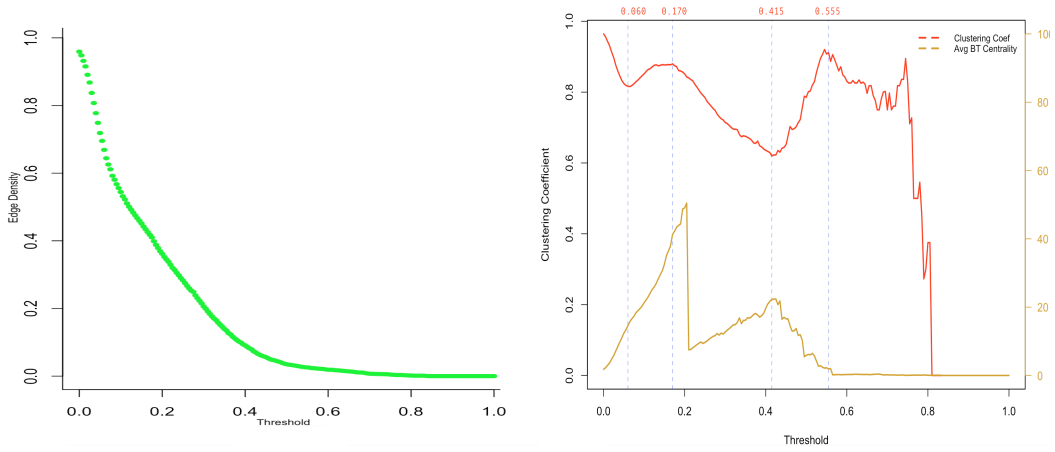


Figure 8 Network Characteristics as functions of different correlation threshold values ranging from 0 to 1. Left panel: Graph Density. Right panel: Clustering coefficient and Average betweenness centrality.

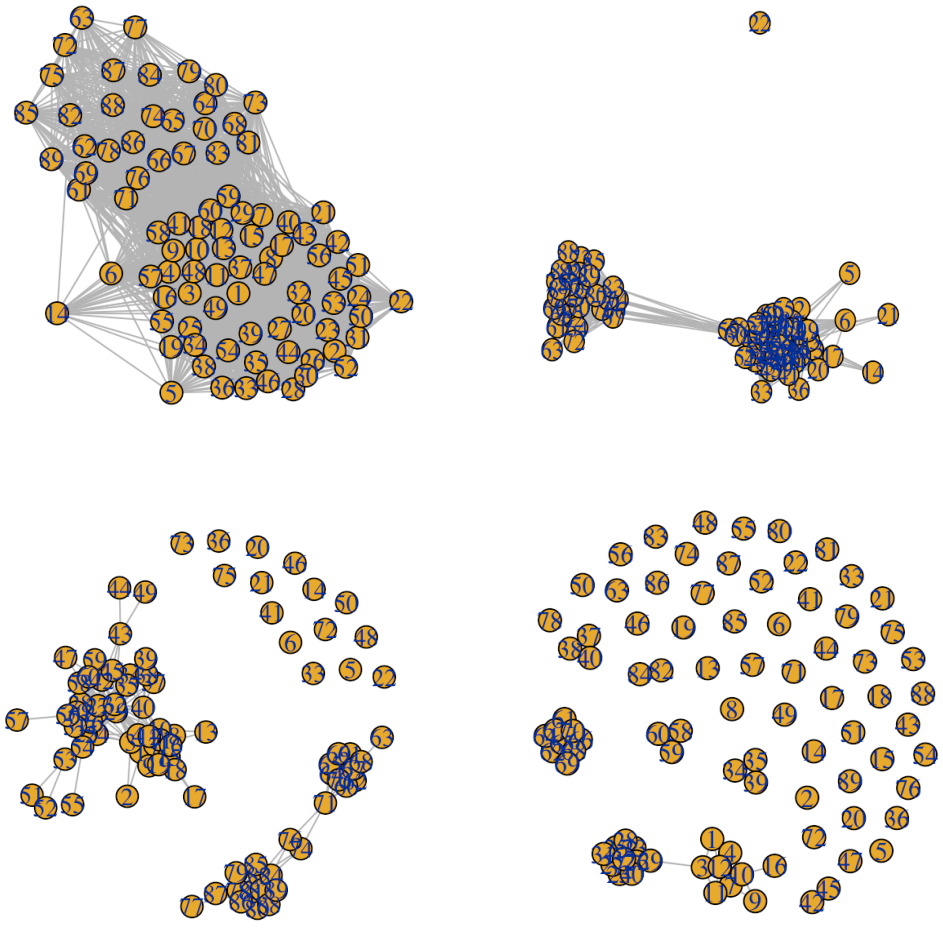


Figure 9 Threshold Network graphs inferred from different correlation thresholds. Top left panel: Network graph at θ=0.06; top right panel: Network graph at θ=0.17; bottom left panel: Network graph at θ=0.415; bottom right panel: Network graph at θ=0.555.

The top left panel of Figure 9. shows a network graph created at threshold θ=0.06. This inferred graph is very dense comparing to other three graphs. One can easily deduce that the threshold value of 0.06 is too small, and cannot produce a meaningful result. The top right panel of Figure 9 displays still very dense graph (inferred at threshold θ=0.17) with one isolated vertex 22 (Biotest Pharmaceuticals Corporation). When θ=0.555, most of vertices are isolated, but majority Chinese banks and US banks are grouped together with other banks in the same county. The bottom left panel of Figure 9 shows a network graph created at threshold θ=0.415 with 89 nodes and 335 edges. This picture exhibits the associations within 89 research corporations more clearly, and hence will be used in the further analysis.

- **Characteristics of the threshold Network**

To understand the significance of the structural properties of the created threshold network, three major network characteristics, vertex degree, betweenness, and local clustering (vertex clustering) are evaluated in this section.

In the left panel of Figure 10, one could see that there are 3 distinct groups in the degree distribution: (1) with vertex degree less than 17, (2) with vertex degree in the range from 19 to 22, and (3) with vertex degree greater than 25. For example, HSBC and TD Bank are two US banks from the third group with 28 degrees implying that these two banks have strong associations (correlation coefficient is greater than 0.415) with other 28 corporations in the network.

The betweenness centrality distribution, illustrated on the middle panel of Figure 10, shows that more than 50 corporations have betweenness equal to zero that means these companies are not passed through in the shortest paths between other pairs of vertices. TD Bank, HSBS has the largest betweenness coefficient 250 and 177, respectively.

The distribution of the local clustering is displayed in the right panel of Figure 10. The local clustering is a measure that is used to describe how likely the neighbors of the target node are likely to form a cluster (neighbors for each other). There are 16 vertices with local clustering equaling to 1, which implies that neighbors connected to these nodes also connect with each other within the neighborhood. The clustering coefficients of TD Bank and HSBC equal to 0.339 and 0.449, respectively.

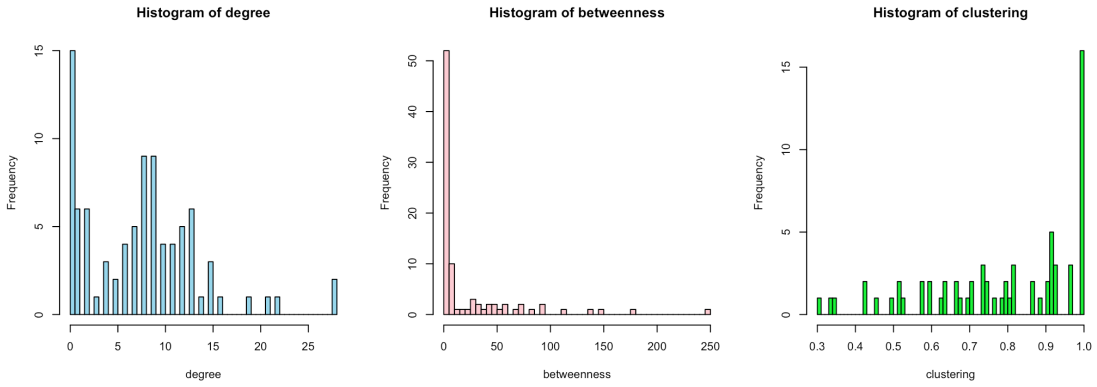


Figure 10 The distributions of different Network Characteristics. Degree distribution (left panel) and Betweenness centrality distribution (central panel), Clustering coefficient distribution (right panel). Figure was drawn at threshold equal to 0.415.

- **Assessing Significance of Network Characteristics**

In order to test the significance of the network characteristics, two random graph simulation techniques are applied in this section. Specifically, the results of 1000 classical random graphs (CRG) and 1000 generalized random graphs (GRG)

simulations are presented here. In the classical random graphs, each graph has the same number of vertices number and the same number of edges as original threshold network G 89 vertices and 335 edges. The difference between CRG and GRG is that the latter one has the required degree sequence, and the previous one does not have this limitation. The results of these two simulation methods are shown in Figure 11 and Figure 12. Figure 11 depicts distributions of the classical random graphs. The two histograms show that both characteristics, namely clustering and betweenness follow a bell shape distribution with means equal to 0.085 and 62.65, respectively. The clustering and betweenness distributions of simulated generalized random graphs are shown in Figure 12 with means equal to 0.208 and 37.88.

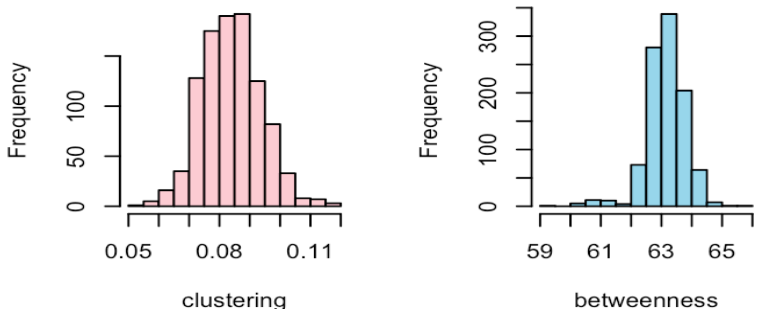


Figure 11 The distribution of classical random graphs' characteristics

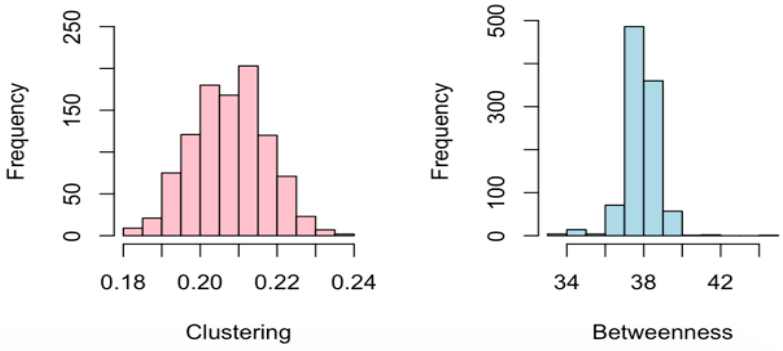


Figure 12 The distribution of generalized random graphs' characteristics

The mean of clustering coefficients in Figure 11 and Figure 12 are both smaller than the statistics values of clustering. From this, one can conclude that in the thresholded network, connected triple nodes are more likely to form triangles, that is the companies prefer to form small cliques. The mean of betweenness in the CRGs and GRGs are greater than the statistics mean value of betweenness. One can conclude there are few companies being very 'important' than other vertices and having centralized position in the thresholded network. In order to test the significance of characteristics in the network, the hypothesis test is applied here:

Ho: $\beta_i = \beta_{i0} \quad , i = average\ degree, density, clustering, average\ betweenness$

Ha: $\beta_i \neq \beta_{i0} \quad , i = average\ degree, density, clustering, average\ betweenness$

The statistics values and corresponding p-values were computed using two random graph simulation methods are presented in Table 5.

	degree	*density*	*clustering*	*Avg. Betweenness*
Statistics	7.528	0.085	0.625	22.25
P-value (Classical Random Graphs)	1.000	1.000	0.000	0.000
P-value (Generalized Random Graphs)	1.000	1.000	0.000	0.000

Table 5 Significance test results of Network characteristics

Two-tailed test is used here to verify the significance of Network characteristics, and one can observe that the obtained p-values of degree and density are both equal to 1 for both classical random graphs and generalized random graphs. This result is expected due to the nature of simulation process of the random graphs. In in both

cases the number of edges (335), and the number of vertices (89) are fixed (the same as in the original graph G), and the density is computed by using formula:

$$den\ (G) = \frac{|E_G|}{\frac{|V_G|(|V_G|-1)}{2}}.$$

Thus, all simulated 1000 random graphs, by the construction, have the same graph density as graph G.

It is worth mentioning that the clustering coefficient is computed by treating the network as integral, at the same time, average betweenness is obtained by averaging 89 vertex betweenness values. The p-value of clustering coefficient and the average betweenness are both zero suggesting the constructed network graph with threshold equaling to 0.415 to be significantly different from the classical (generalized) random graph and to capture the important associations in the dataset.

In order to make the network graph more visually understandable, the decorated network will be presented in the next section.

5.1.2 Visualization of Network

Decorating graph layout can be very helpful in visualizing large network. Here, three different vertex shapes, namely circle, square and triangle, are used to denote the three countries: Germany, US, and China. Four different colors of vertex, red, green, blue, and purple are utilized to represent the four industrial sectors including: Banking, Manufacturing, Telecommunication, and Pharmaceutical, respectively. Comparing the decorated graph in Figure 13 to the non-decorated in Figure 9, one can observe that the decorated graph is more readable and it gives a more general overall impression of the relationships between corporations across different countries and sectors. For

example, Figure 13 shows that Chinese companies (triangles) are close to each other and do not have strong association with US or German companies. The US corporations (circles) and German companies (squares) are in a one connected group, and the US banks locate in the central of the group. As well, three Chinese telecommunications (blue triangles) trading in the US are grouped with majority US and German companies instead of Chinese companies.

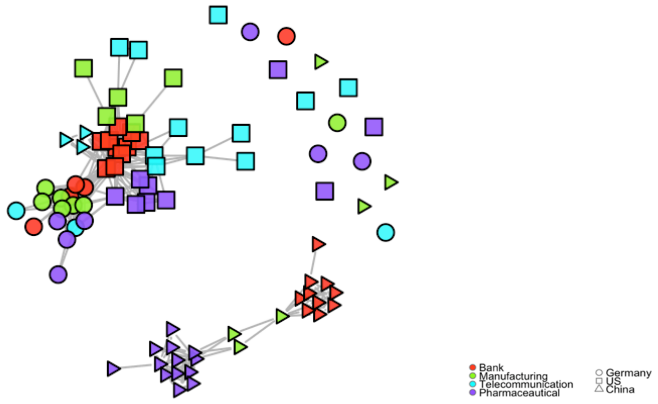


Figure 13 Network Visualization

5.2 Network Community detection

One of the research goals of this thesis is to identify groups of companies that exhibit the most similar stock market trends. Here, hierarchical clustering method and reduced network technique are utilized to detect the associations between corporations and the relationship within and between different industries, respectively.

In order to present the graph in a readable way, the company ID are created. The country initial and sector initial are utilized here. For example, the first letters of the three research countries Germany, US, and China are shortened as G, U and C, respectively. Similarly, the four research industrial sectors, Banking, Pharmaceuticals,

Manufacturing, and Telecommunications are shortened as character B, P, M, and T. As well, companies in the same country and sectors are labeled using the short name of belonging sectors and numbers. For example, there are 3 Chinese telecommunication companies, thus, these three are labeled as 'CT1','CT2','CT3'.

Hierarchical clustering dendrogram is shown in Figure 14 with 22 clusters. If one cuts the dendrogram tree at height 1.48, the graph will be divided into five main communities and each community will have more than 4 components. In the Figure 14, one can see that the first community consists by ten strongly associated Chinese banks; and nine Chinese pharmaceutical corporations form the second community. Chinese Manufacturing companies have strong relationship with Chinese pharmaceutical sector, but they are not clustered in one group when the height =1.48. There are total of 8 US banks that contribute to the third community. The fourth community combines four leading German banks and 5 German manufacturing corporations. The last cluster includes the majority of German companies, US companies and 3 Chinese telecommunications trading in the US.

Figure 14 shows that 3 Chinese telecommunication companies trading in the US (with ID CT1, CT2 and CT3) have the high correlation with US companies. German sectors, especially Auto-Manufacturing (with ID GM), also have the strong association with the companies in all four US industries. Except for telecommunication, all Chinese companies have strong relationships within the country. The US and German sectors have strong associations. Five detected clusters can be described as follows: companies that correspond to Chinese government

control, Chinese non-government owned companies, US Banks, manufacturing and banks in Germany, and the other US and German companies.

In this section, the clustering method detected the companies that exhibit the most similar stock market trends. In the next section, the reduced network is used to evaluate associations between different industries.

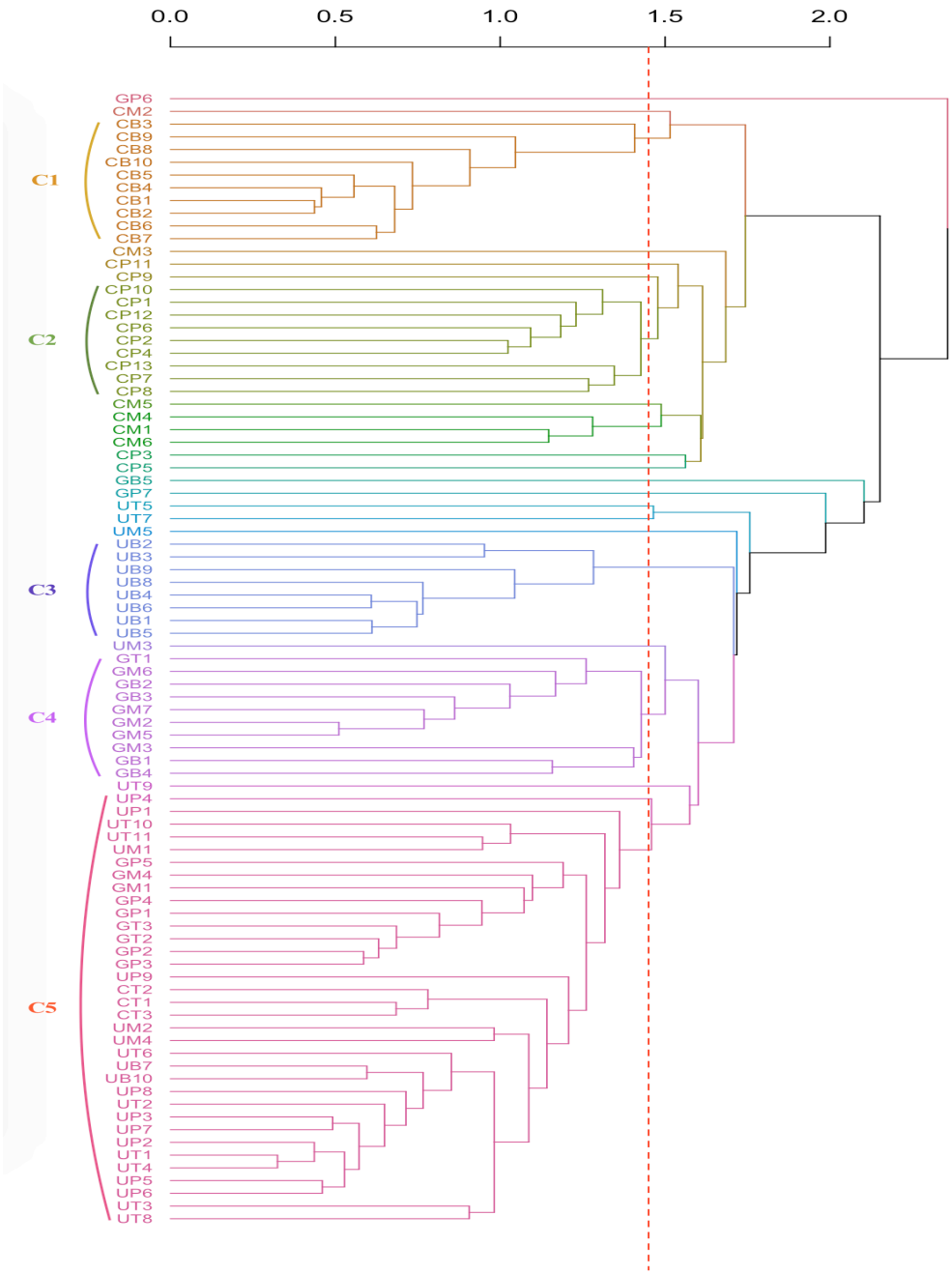


Figure 14 Hierarchical clustering dendrogram

- **Reduced Network**

To evaluate associations between different industries, a reduced network graph is created, where: (1) vertex size is determined by the average of inner correlations (correlations between companies within the same country and industry), and (2) width of graph edges is defined by the average correlation between companies in different industries.

The first letters of the three research countries are used here as the short names, for example, G, U and C are used to denote Germany, US, and China, respectively. Similarly, the character B, P, M, and T are used to represent four research industrial sectors, Banking, Pharmaceuticals, Manufacturing, and Telecommunications.

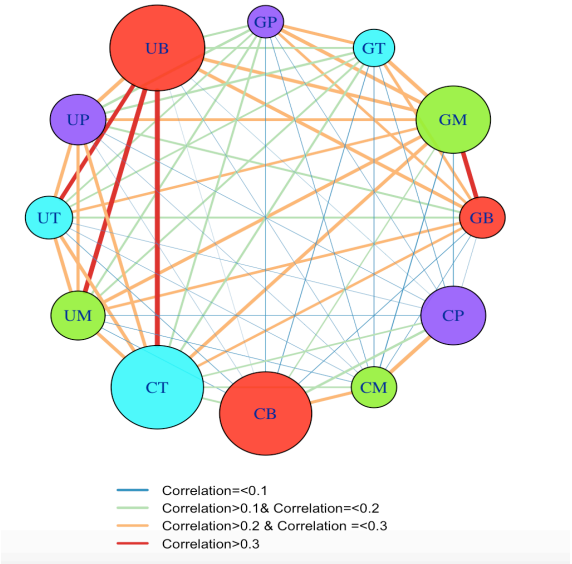


Figure 15 The reduced network graphs

Figure 15 shows that Chinese banks and Telecommunication corporations have strong inner correlations and therefore visualized with a large vertex sizes. The US

banking and German manufacturing also have big size of vertices compared to other US and German sectors. The strong associations between the German companies and US companies are visualized by thick edges in the graph in Figure 15. All Chinese companies have strong associations within the country.

5.3 Dynamitic Analysis Result

To explore the annual changes in company/country/industry associations in the past seven years (from 2009 to 2015), the annual dynamic association networks with the spanning trees are created. From Table 6 and Figure 16, one could find that before 2009 European Debt Crisis, most German Pharmaceutical corporations had almost no relationship with other companies, but after crisis, the ice was broken and stock returns of many pharmaceutical corporations followed the same trends as German and US Banks.

Characteristics / *Year*	*Density*	*Clustering Coefficient*	*Avg. of Betweenness coefficient*
2009	0.122	0.670	16.860
2010	0.175	0.689	12.79
2011	0.283	0.857	8.494
2012	0.111	0.643	11.148
2013	0.063	0.705	24.932
2014	0.056	0.763	15.584
2015	0.153	0.770	19.404

Table 6 Table of Network Characteristics across 7 research years from 2009 to 2015.

The network characteristics are summarized in Table 6, and include graph density, clustering coefficient and average betweenness. The network graph became increasingly denser and started to form more connected clusters in the early recovery period, 2-3 years after the crisis (2009 to 2011)

In order to detect the annual changes in company/country/industry association for some specific companies and sectors, the annual dynamic spanning trees are constructed and presented in Figure 16. As introduced in the Chapter 4, the spanning tree is a subgraph of the network graph that simply connects all originally connected nodes and omits the loops. One can see that most Chinese companies maintained their connections and did not change them much in the time period from 2009 to 2014. However, the associations within Chinese companies became denser in 2015 due to the Chinese stock market crash. The US pharmaceutical companies had weak associations during the crisis, and stronger inner relationships after the crisis.

Discovering the associations between different companies statically and dynamically from 2009 to 2015 is one of the research goals of this thesis, which has been achieved in this section. On the other hand, exploring the relationship between the accuracy of classification models and network node properties is the last research goal of this thesis. In the next section, four classification models will be used to predict the future stock movement and discover the associations between accuracy rate and node properties graphically.

Figure 16 Annually Dynamic Network in the time period from 2009 to 2017

5.4 The Performance of Four Classification Models

Forecasting the future return trend movement of different companies has been particularly popular in the field of financial data analysis. One of goals of this thesis is to discover if there is a relationship between the accuracy of classification of stock return movements and network node properties.

Here, four classification methods, namely LDA, QDA, KNN, and Logistic regression are employed to predict the future return movements of 89 companies. The predictive accuracy rate is used to evaluate the performance of different models.

In the research data, first six-year stock returns (from 2009 to 2014) are used as training data and last year research data (in 2015) are utilized as test set. On the other hand, the dependent variable (Y) is a two-level categorical variable that represents movement directions of stock price returns; the predictors are the stock returns of the eighty-nine research companies. Formally, the previous one day stock returns were utilized, ($X_{1,t-1}$, $X_{2,t-1}$, $X_{3,t-1}$, $X_{4,t-1,\ldots}$, $X_{p,t-1}$), to class the next day stock movement direction of a target company $Y_{t.}$

- **Model performance across different countries**

A typical approach to assessing model performance is separating the data into two parts, training set and test set. The training data is used to build a model and estimate the related parameters; the test data is normally used to test the performance of the developed model. To evaluate the model performance and select the better performance classification model, the true accuracy rates are applied in this section. The true accuracy rate of four classification models across three countries are listed in the following four tables, Table 7- Table 10. Table 7, for example, suggests that the mean accuracy of LDA model for 32 Chinese companies is slightly greater than 0.5 and the mean accuracy of 35 US companies is smaller than 0.5 which is even worse than the random guess. The overall performance of LDA model is not ideal in prediction movement of price return movements.

SUMMARIZATION	GERMANY	US	CHINA
N	22	35	32
MIN	0.478	0.431	0.474
MEDIAN	0.530	0.504	0.538
MEAN	0.528	0.498	0.537
MAX	0.599	0.551	0.590
SD	0.035	0.032	0.033

Table 7 The prediction accuracy of LDA model across 3 different countries

The QDA model performance across three different countries is displayed in Table 8. The mean prediction accuracy is 0.5, as low as a random guess. The maximum accuracy among US companies is 0.59, which is slightly better than a random guess. It can be interpreted as follows. If a company price return is predicted to 'increase' tomorrow by using QDA model, then, there will be 59% chance this company would actually increase tomorrow.

SUMMARIZATION	GERMANY	US	CHINA
N	22	35	32
MIN	0.413	0.431	0.444
MEDIAN	0.493	0.500	0.502
MEAN	0.492	0.503	0.506
MAX	0.530	0.590	0.569
SD	0.027	0.034	0.030

Table 8 The prediction accuracy of QDA model across 3 different countries

The closest 19 neighbors are used here to create the KNN predictor model. In order to choose the proper neighbor size K, the general rule is applied here, where K is one half of the square root of the samples size in the training dataset. The formula can be written as $K = \sqrt{\frac{N}{2}}$. If there are 1388 observations in the training dataset, then K=19.

SUMMARIZATION	GERMANY	US	CHINA
N	22	35	32
MIN	0.469	0.452	0.465
MEDIAN	0.504	0.504	0.528
MEAN	0.511	0.502	0.527
MAX	0.577	0.594	0.577
SD	0.033	0.027	0.026

Table 9 The prediction accuracy of KNN model across 3 different countries

Table 9 shows the average accuracy of KNN model to be 0.51. The predictions provided by KNN for Chinese companies are slightly higher than the predictions for other two countries, but the differences are still minor.

The classification accuracy results for logistic regression are listed in Table 10. One can find that the average accuracy rates of German and Chinese companies are around 0.53, greater than accuracy rates achieved by previously described three classification models. The mean accuracy rate of the US companies is slightly higher than a random guess.

SUMMARIZATION	GERMANY	US	CHINA
N	22	35	32
MIN	0.482	0.418	0.482
MEDIAN	0.534	0.504	0.5323
MEAN	0.529	0.505	0.530
MAX	0.599	0.560	0.5905
SD	0.033	0.035	0.0337

Table 10 The prediction accuracy of Logistic regression model across different countries

To compare performance across three countries for four outlined models, four parallel box charts are created below. The first three panels of Figure 17 compare the performance of the models in each country separately. The last panel of Figure 17 (bottom right) is created by combining all companies in all three countries and

compares the accuracy of the four classification models. The top left panel of Figure 17 shows that for German companies the QDA model has the poorest performance, LDA and Logistic regression model have a better prediction than other two models. In the US, all four model medians are around 0.5. The LDA and Logistic regression models perform better than the other two models for the selected Chinese corporations. The last chart suggests that the performance of LDA and Logistic models are similar, but the logistic regression model has a slightly higher standard division than LDA.

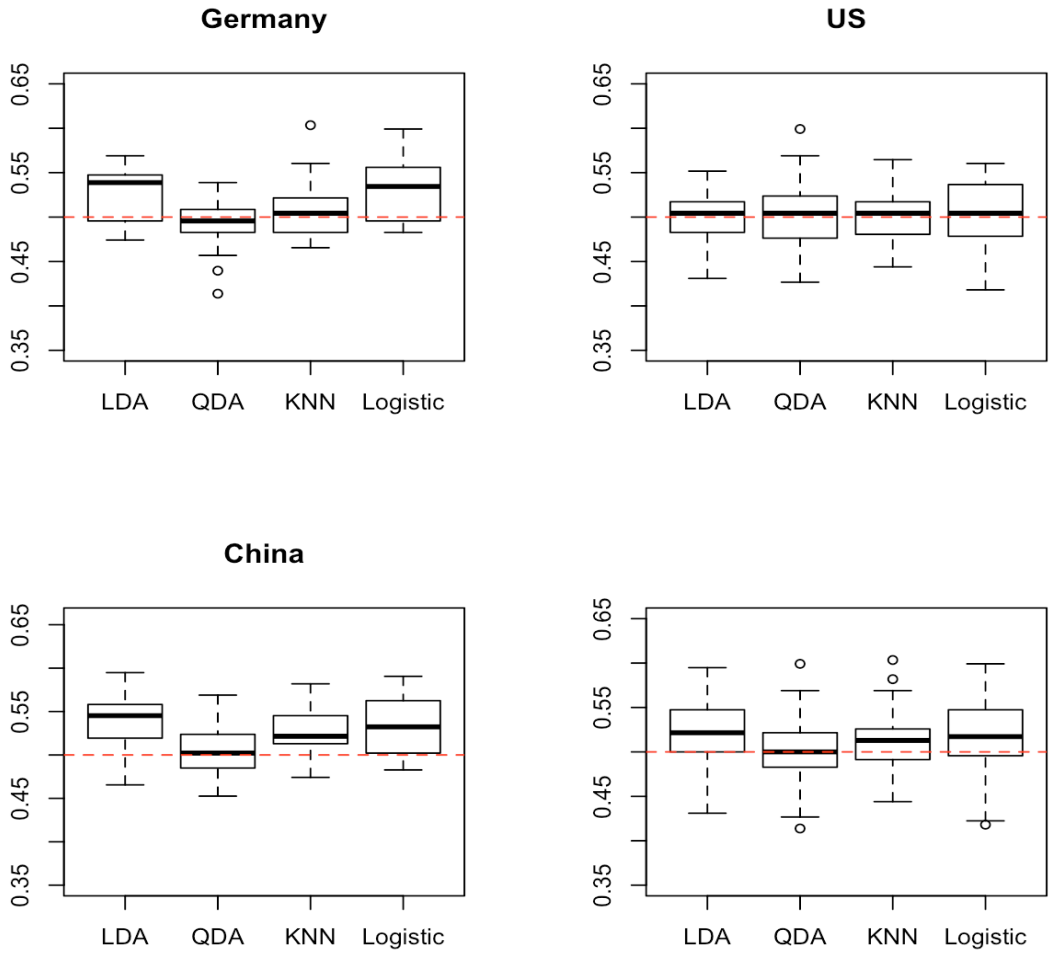


Figure 17 The predictive performance of 4 classification models. Top left panel: compare model performance across Germany. Top right panel: compare model

performance across US. Bottom left panel: compare model performance across China. Bottom right panel: compare model performance in general

5.5 Associations Between Network Features and Regression Model Performance

The accuracy rates of four classification models were compared in the previous section, and the LDA and Logistic regression model had a better performance than the other two selected models. The aim of his section is to explore the association between Logistic regression classification accuracy and threshold network properties.

The main reason of using Logistic regression model instead of LDA is as follows: the LDA model assumes the predictors to follow a multivariate normal distribution; however, the logistic regression does not carry normal distribution requirement for predictors. Unfortunately, the previous preliminary results showed that the research data returns were not normally distributed. Thus, the results of LDA are suspect, and this research will use the logistic regression as a more reliable model to do the further analysis.

Here, in this section, two graphical methods, the scatter chart and the level plot are applied to detect the relationship between logistic regression model and node properties. The scatter plots are displayed in Figure 18, and the level plots are presented in Figure 19. It is worth noting that the nodes in the network are research companies and the accuracy logistic regression classification rates are calculated for each company separately.

From Figure 18, one can clearly see that vertex clustering has a positive relationship with the classification accuracy rate. The maximum accuracy is equal to 0.6 when the vertex clustering equaling to 1, where neighbors of the node also connected and formed a clique. When the degree of vertex locates between 8 and 13,

the accuracy rates of most companies are greater than 0.5 and deviation of accuracies has a small variability. The betweenness centrality of the most vertices is less than 100, and there are 9 out of 89 corporations that are outliers and have incredibly high betweenness. The accuracy of the nine organizations varies in a large range and has a great variability.

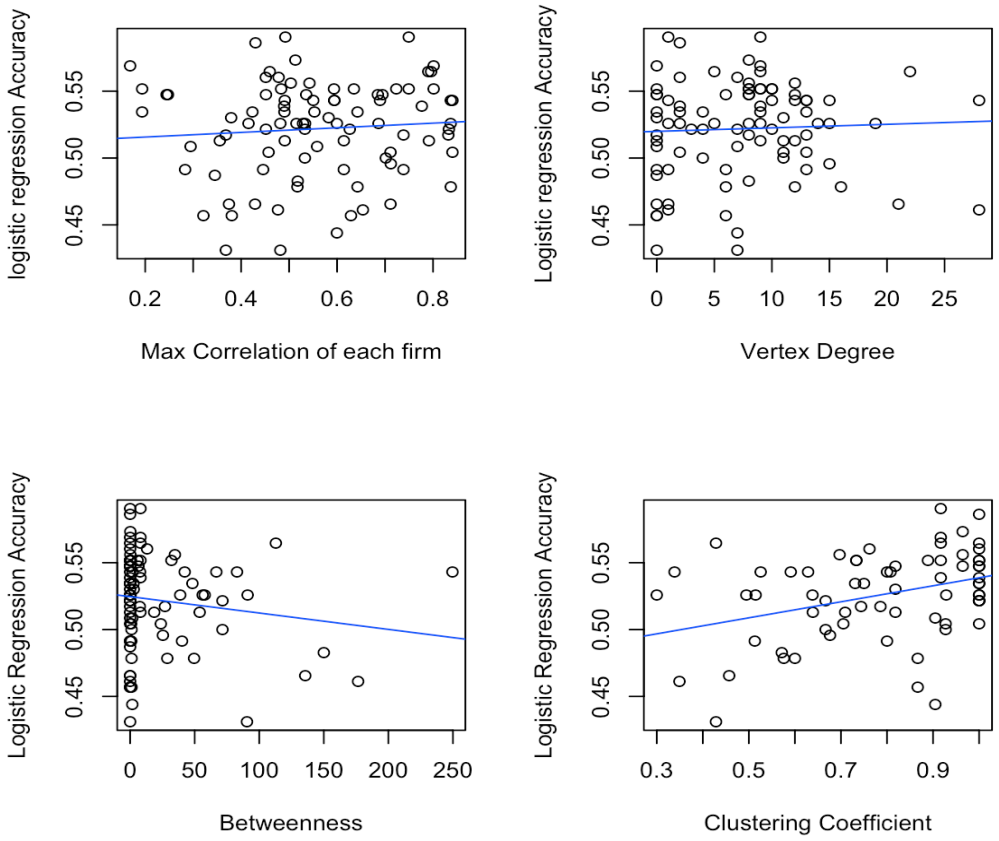


Figure 18 The scatter plots between classification accuracy rates and threshold network node properties.

The scatter plot is a tool that detects (if present) the relationships between two variables. Here, the level plot is applied to detect the association between 3 variables. The main idea of level plot is that the X-axis and Y-axis are divided into different cells. If there are many observations located in the same cell, then, this cell will be divided

into the smaller cells. Thus, the big size cell does not stand for most case happening in there, on opposite, it stands the event barely happened in that coordinate. In addition, the shade of color can represent the value of accuracy rate.

The level plot in Figure 19 describes the relationships between the network node properties and the model accuracy rate, where the X-axis and Y-axis are the network characteristics, and different shades of green and red color denote if the accuracy rate are greater than 0.5.

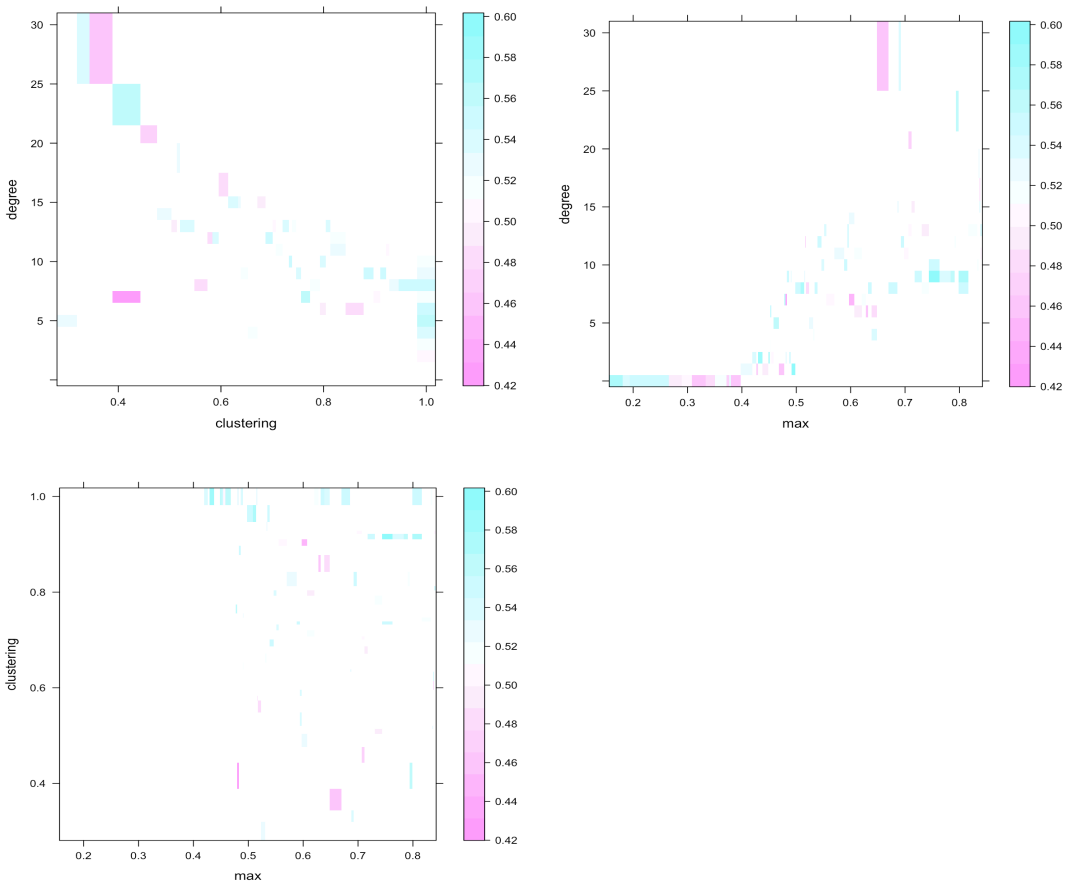


Figure 19 The level plots of the relationship between classification accuracy rate and network node properties. Left top panel: accuracy rate versus vertex clustering and vertex degree. Right top panel: accuracy rate versus vertex maximum correlation

coefficient with other companies and vertex degree. Left bottom panel: accuracy rate versus vertex maximum correlation and vertex clustering

From the top left panel of Figure 19, one could easily see that the cells color are bright blue when vertex has clustering coefficient being greater than 0.7 and degree being bigger than 7. One also can note from this figure that vertex degree and clustering has a negative relationship, as the vertex degree decreases, the clustering coefficient increases. The 'important' company having extremely big degree is difficult to form a small clique. The top right panel of Figure 19 shows the maximum company correlation coefficient has positive correlation with vertex degree. There is no clear pattern in the bottom left panel of Figure 19.

Hence, based on the observed results from the scatter plots and the level plots one can conclude that the movement direction of stock price return would be easier to classify by using logistic regression model than other companies, if the vertex satisfies the following conditions: (1) vertex are more like to be a follower instead of a leader with eight to thirteen neighbors in the association network, and (2) vertex prefers to form small cliques, where its connected neighbors are also close and connected.

CHAPTER 6

CONCLUSION

In order to reach the five goals of this thesis, we used the 89 companies daily stock price data collected from a publicly available source, Yahoo Finance, for a time period from 2009 to 2015. To reduce the variance of the data, the close price data was converted into daily returns, and then used to compute a correlation matrix and create a corresponding association network.

After obtaining the association network, the community detection method, agglomerative hierarchical clustering, was applied in this study to identify companies that exhibit the most similar return trends. The results suggested that the companies that traded in the same stock market and/or belonged to the same industrial sectors had significant associations. Specifically, the Chinese companies had higher inner correlations in banking and telecommunication sectors; while the US and the German companies had stronger associations in banking and auto-manufacturing sectors.

In addition to detecting static associations between companies over the research years from 2009 to 2015, the annually dynamic networks were created to assess annual changes in associations between selected companies during a special financial period, i.e.2009 European Debt Crisis. The results showed that the associations among companies became stronger and more companies tended to be grouped together in the network during European Debt Crisis and in the early recovery periods.

Another focus of this thesis was on discovering the relationship between classification accuracy rates and the network node properties. Four classification

models, namely Linear Discriminant Analysis, Quadratic Discriminant Analysis, K-Nearest Neighbors, and Logistic Regression were created and evaluated. The results revealed the superior performance of the logistic regression method compared to the other three classification methods, particularly for the Chinese companies. Thus, the logistic regression was utilized later to detect the relationship between model accuracy rates and network node properties. Two graphical tools were applied in this thesis. The results illustrated that companies that acted as followers and belonged to medium-size clusters with eight to thirteen neighbors in the association network were easier to classify than the other companies.

Even though the logistic regression had a better performance comparing to the rest three classification methods, LDA, QDA and KNN, especially for Chinese corporations. However, it is worth to mention that the accuracy of Logistic regression model of Chinese companies has mean=0.530 and standard deviation=0.033. If one assumes the model accuracy follows a normal distribution, then the 95% confident interval for the true accuracy will be in the range of from 0.478 to 0.598. Based on this result, it is difficult to conclude that the logistic regression has a statistically significant performance compared to a random guess.

Thus, in order to improve the classification accuracy, in a future study, I aim to use the data fusion classification method proposed by Dr. Natallia Katenka to predict future price movement more accurately. This method will leverage the vertex own historic data and the related neighbors data as predictors to forecast the future return directions.

APPENDIX

Filter

	date	Deutsche.Bank.stock	Commerzbank.AG..Frankfurt.M.	X.Postbank.AG	X.Aareal.Bank.AG	X.IKB.Deutsche.Industriebank.AG	AUDI	BMW	Porsche
1	2009-01-05	-0.0436	0.0057	-0.0507	-0.0387	-0.0289	-0.0942	-0.0226	-0.0238
2	2009-01-06	-0.0428	0.0378	-0.0135	0.0177	-0.0558	0.0061	0.0182	0.0481
3	2009-01-07	-0.1438	-0.0874	-0.0543	-0.0103	0.0443	0.0182	-0.0017	0.0046
4	2009-01-08	-0.1185	-0.0634	-0.0609	-0.0098	-0.0496	-0.0014	0.0207	-0.0407
5	2009-01-09	-0.0403	-0.0621	-0.0086	-0.0809	-0.0186	-0.0164	-0.0227	-0.0164
6	2009-01-12	-0.0645	0.1012	-0.0184	-0.0666	-0.1351	-0.0092	-0.0557	-0.0092
7	2009-01-13	-0.1175	-0.1963	-0.0992	-0.0241	-0.0548	-0.0048	-0.0612	-0.0048
8	2009-01-14	-0.1185	-0.2121	-0.0665	-0.0382	-0.0175	-0.0046	-0.0286	-0.1524
9	2009-01-15	-0.0758	-0.1014	-0.0304	0.0159	0.0201	0.0842	0.0036	-0.0031
10	2009-01-16	-0.0183	-0.1403	-0.1096	-0.2755	0.0125	0.0023	-0.0222	-0.0408
11	2009-01-20	-0.0205	0.0942	0.0335	-0.0101	-0.0062	-0.0049	0.0083	-0.0426
12	2009-01-21	0.0255	0.0077	0.0090	-0.1025	0.0392	0.0064	-0.0195	-0.0227
13	2009-01-22	0.0153	-0.0465	-0.0540	-0.0490	-0.0611	-0.0068	-0.0452	0.0427
14	2009-01-23	0.0603	0.0831	0.0754	0.0748	0.0672	0.0064	0.0399	0.0324
15	2009-02-02	0.0239	0.0024	0.0722	0.0091	-0.0108	0.0091	0.0379	-0.0245
16	2009-02-03	0.0103	0.0665	0.0348	0.0315	0.0694	0.0045	0.0709	0.0247
17	2009-02-04	-0.0126	-0.0242	-0.0510	-0.0244	-0.0678	-0.0082	0.0033	0.0033
18	2009-02-05	0.0476	0.0274	0.0546	0.0568	-0.0268	-0.0011	0.1085	-0.0495
19	2009-02-06	0.0935	-0.0031	0.0716	0.0269	0.0266	-0.2062	0.0583	0.0825
20	2009-02-09	-0.0298	-0.0016	-0.0328	-0.0581	0.0084	-0.0037	-0.0302	0.0007
21	2009-02-10	-0.0281	0.0051	-0.0001	-0.0364	0.0251	-0.0012	0.0121	-0.0012
22	2009-02-11	-0.0482	-0.0239	0.0036	0.0044	-0.0571	0.0214	0.0002	-0.0038
23	2009-02-12	-0.0868	-0.0096	0.0202	0.0372	0.0812	0.0023	0.0460	-0.0617
24	2009-02-17	0.0597	0.0235	-0.0002	-0.0231	-0.0447	-0.0055	-0.0213	-0.0212
25	2009-02-18	0.0334	0.0605	0.0101	-0.0141	0.0052	0.0039	0.0232	0.0039
26	2009-02-19	-0.0876	-0.0267	-0.0990	-0.0139	0.0396	-0.0158	-0.0301	-0.1042
27	2009-02-20	-0.0165	-0.0211	-0.0386	0.0270	0.0011	0.0140	-0.0420	-0.0745

Figure 1 Data sample.

Name	Symbol	ID	Cl	Country	Industry	Name	Symbol	ID	Cl	Country	Industry
Commerzbank AG, Frankfurt/M.	CBK.DE	GB1	GB	Germany	Bank	T-mobile US	TMUS	UT1	UT	US	Telecommunication
Postbank AG	DPB.DE	GB2	GB	Germany	Bank	U.S. Cellular	USM	UT2	UT	US	Telecommunication
Deutsche Bank stock	DBK.DE	GB3	GB	Germany	Bank	Sprint Corp	S	UT3	UT	US	Telecommunication
Aareal Bank AG	ARL.DE	GB4	GB	Germany	Bank	Windstream Hldgs	WIN	UT4	UT	US	Telecommunication
IKB Deutsche Industriebank AG	IKB.F	GB5	GB	Germany	Bank	FairPoint	FRP	UT5	UT	US	Telecommunication
AUDI	NSU.DE	GM1	GM	Germany	Manufacturing	Cincinnati Bell Inc.	cbb	UT6	UT	US	Telecommunication
BMW	BMW.DE	GM2	GM	Germany	Manufacturing	Hawaiian Telcom	HCOM	UT7	UT	US	Telecommunication
Porsche	POAHF	GM3	GM	Germany	Manufacturing	General Motors Company	GM	UM1	UM	US	Manufacturing
Volkswagen	VOW3.DE	GM4	GM	Germany	Manufacturing	Ford-Werke GmbH	F	UM2	UM	US	Manufacturing
Mercedes-Benz	DAI.DE	GM5	GM	Germany	Manufacturing	chrysler	FCAU	UM3	UM	US	Manufacturing
Continental	CON.DE	GM6	GM	Germany	Manufacturing	Polaris	PII	UM4	UM	US	Manufacturing
ThyssenKrupp AG	TKA.DE	GM7	GM	Germany	Manufacturing	Tesla Motors	TSLA	UM5	UM	US	Manufacturing
Drillisch Aktiengesellschaft	DRI.DE	GT	GT	Germany	Telecommunicati	China Mobile Ltd	CHL	CT	CT	China	Telecommunication
Telef„nica Germany-O2	O2D.DE	GT	GT	Germany	Telecommunicati	China Unicom	CHU	CT	CT	China	Telecommunication
Deutsche Telekom	DTE.DE	GT	GT	Germany	Telecommunicati	China Telecom Corporation Limited	CHA	CT	CT	China	Telecommunication
Bayer AG	BAYN.DE	GP1	GP	Germany	Pharmaceautical	ICBC	601398.SS	CB1	CB	China	Bank
Fresenius Medical Care	FME.DE	GP2	GP	Germany	Pharmaceautical	ChinaConstructionBankget	601939.SS	CB2	CB	China	Bank
Beiersdorf	BEI.DE	GP3	GP	Germany	Pharmaceautical	AgriculturalBankofChina	601288.SS	CB3	CB	China	Bank
Stada-Arzneimittel AG	SAZ.DE	GP4	GP	Germany	Pharmaceautical	BankofChina	601988.SS	CB4	CB	China	Bank
Merck KGaA	MRK.DE	GP5	GP	Germany	Pharmaceautical	BankofCommunications	601328.SS	CB5	CB	China	Bank
Paion AG	PA8.DE	GP6	GP	Germany	Pharmaceautical	ChinaMerchantsBank	600036.SS	CB6	CB	China	Bank
Biotest Pharmaceuticals Corporation	BIO.DE	GP7	GP	Germany	Pharmaceautical	MinShengBank	600016.SS	CB7	CB	China	Bank
JPMorgan Chase	JPM	UB1	UB	US	Bank	ShanghaiPudong	600000.SS	CB8	CB	China	Bank
Bank of America Corp	BAC	UB2	UB	US	Bank	IndustrialBank	601166.SS	CB9	CB	China	Bank
CITIGROUP	C	UB3	UB	US	Bank	ChinaCiticBank	601998.SS	CB10	CB	China	Bank
WELLS FARGO & COMPANY	WFC	UB4	UB	US	Bank	SAIC Motor	600104.SS	CM1	CM	China	Manufacturing
BANK OF NEW YORK MELLON CORPO	BK	UB5	UB	US	Bank	Great Wall Motor	601633.SS	CM2	CM	China	Manufacturing
U.S.BANCORP	USB	UB6	UB	US	Bank	Beiqi Foton Motor Co. Ltd.	600166.SS	CM3	CM	China	Manufacturing
HSBC North American Holdings Inc.	HSBC	UB7	UB	US	Bank	China Grand Automotive Services Co. L	600297.SS	CM4	CM	China	Manufacturing
PNC Financial Services Group Inc	PNC	UB8	UB	US	Bank	Sinomach Auto	600335.SS	CM5	CM	China	Manufacturing
Capital One Financial Corp	COF	UB9	UB	US	Bank	Anhui Jianghuai Automobile	600418.SS	CM6	CM	China	Manufacturing
TD BANK US HOLDING COMPANY	TD	UB10	UB	US	Bank	Yunnan Baiyao Group	000538.SZ	CP1	CP	China	Pharmaceautical
Gilead Sciences	GILD	UP1	UP	US	Pharmaceautical	Shanghai Pharmaceuticals Holding Co L	601607.SS	CP2	CP	China	Pharmaceautical
Johnson.Johnson	JNJ	UP2	UP	US	Pharmaceautical	China Meheco	600056.SS	CP3	CP	China	Pharmaceautical
Pfizer	PFE	UP3	UP	US	Pharmaceautical	Harbin Pharmaceutical Group Co. Ltd.	600664.SS	CP4	CP	China	Pharmaceautical
Roche	RHHBY	UP4	UP	US	Pharmaceautical	Nanjing Pharmaceutical Co., Ltd	600713.SS	CP5	CP	China	Pharmaceautical
Novartis	NVS	UP5	UP	US	Pharmaceautical	North China Pharmaceutical Co., Ltd	600812.SS	CP6	CP	China	Pharmaceautical
GlaxoSmithKline	GSK	UP6	UP	US	Pharmaceautical	Chongqing Taiji Industry (Group) Co., Lt	600129.SS	CP7	CP	China	Pharmaceautical
Merck	MRK	UP7	UP	US	Pharmaceautical	Shandong Lukang Pharmaceutical Co., L	600789.SS	CP8	CP	China	Pharmaceautical
Sanofi	SNY	UP8	UP	US	Pharmaceautical	Tianjin Zhong Xin Pharmaceutical Grou	600329.SS	CP9	CP	China	Pharmaceautical
AstraZeneca	AZN	UP9	UP	US	Pharmaceautical	Guangzhou Baiyunshan Pharmaceutica	600332.SS	CP10	CP	China	Pharmaceautical
ATT	T	UT1	UT	US	Telecommunicati	TongRenTang	600085.SS	CP11	CP	China	Pharmaceautical
CenturyLink Inc	CTL	UT2	UT	US	Telecommunicati	Zhejiang Hisun	600267.SS	CP12	CP	China	Pharmaceautical
Frontier	FTR	UT3	UT	US	Telecommunicati	Joincare Pharmaceutical Group	600380.SS	CP13	CP	China	Pharmaceautical
Verizon	VZ	UT4	UT	US	Telecommunication						

Figure 2 Name, country and industrial sectors of selected companies.

BIBLIOGRAPHY

Alrasheedi, Melfi. 2012. 'Predicting Up/Down Direction using Linear Discriminant Analysis and Logit Model: The Case of SABIC Price Index', *Research Journal of Business Management*, 6.

Banik, S., A. F. Khodadad Khan, and M. Anwer. 2014. 'Hybrid machine learning technique for forecasting Dhaka stock market timing decisions', *Comput Intell Neurosci*, 2014: 318524.

Benjamini, Y., & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society*. *Series B (Methodological)*, 289-300.

Bookstaber, Richard. 2007. 'A demon of our own design: Markets, hedge funds, and the perils of financial innovation'.

Dai, Y., D. Han, and W. Dai. 2014. 'Modeling and computing of stock index forecasting based on neural network and Markov chain', *ScientificWorldJournal*, 2014: 124523.

Ehrman, Douglas S. 2006. *The handbook of pairs trading: strategies using equities, options, and futures* (John Wiley & Sons).

Heimo, Tapio, Jari Saramäki, Jukka-Pekka Onnela, and Kimmo Kaski. 2007. 'Spectral and network methods in the analysis of correlation matrices of stock returns', *Physica A: Statistical Mechanics and its Applications*, 383: 147-51.

Hotelling, H. 1953. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological),* 15(2), 193-232.

Huang, Wei, Yoshiteru Nakamori, and Shou-Yang Wang. 2005. 'Forecasting stock market movement direction with support vector machine', *Computers & Operations Research*, 32: 2513-22.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning* (Springer).

Kara, Yakup, Melek Acar Boyacioglu, and Ömer Kaan Baykan. 2011. 'Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange', *Expert Systems with Applications*, 38: 5311-19.

Leung, Mark T., Hazem Daouk, and An-Sing Chen. 2000. 'Forecasting stock indices: a comparison of classification and level estimation models', *International Journal of Forecasting*, 16: 173-90.

Lim, Kyuseong, Min Jae Kim, Sehyun Kim, and Soo Yong Kim. 2014. 'Statistical properties of the stock and credit market: RMT and network topology', *Physica A: Statistical Mechanics and its Applications*, 407: 66-75.

Narayan, Paresh Kumar, and Deepa Bannigidadmath. 'Does Financial News Predict Stock Returns? New Evidence from Islamic and Non-Islamic Stocks', *Pacific-Basin Finance Journal*.

Nguyen, Thien Hai, Kiyoaki Shirai, and Julien Velcin. 2015. 'Sentiment analysis on social media for stock movement prediction', *Expert Systems with Applications*, 42: 9603-11.

Nobi, Ashadun, Sungmin Lee, Doo Hwan Kim, and Jae Woo Lee. 2014. 'Correlation and network topologies in global and local stock indices', *Physics Letters A*, 378: 2482-89.

Peng, Yangtuo ; Jiang, Hui. 2015. 'Leverage Financial News to Predict Stock Price Movements Using Word Embeddings and Deep Neural Networks'.

Sienkiewicz, Adam, Tomasz Gubiec, Ryszard Kutner, and Zbigniew R Struzik. 2013. 'Dynamic structural and topological phase transitions on the Warsaw Stock Exchange: A phenomenological approach', *arXiv preprint arXiv:1301.6506*.

Song, Dong-Ming, Michele Tumminello, Wei-Xing Zhou, and Rosario N Mantegna. 2011. 'Evolution of worldwide stock markets, correlation structure, and correlation-based graphs', *Physical Review E*, 84: 026108.

Wang, Shuai, and Wei Shang. 2014. 'Forecasting Direction of China Security Index 300 Movement with Least Squares Support Vector Machine', *Procedia Computer Science*, 31: 869-74.

WilmerHale. 2016. '2016 IPO Report', WilmerHale. http://www.wilmerhale.com/uploadedFiles/Shared_Content/Editorial/Publications/Documents/2016-WilmerHale-IPO-Report.pdf.

Zhang, Wenping, Chunping Li, Yunming Ye, Wenjie Li, and Eric WT Ngai. 2015. 'Dynamic business network analysis for correlated stock price movement prediction', *IEEE Intelligent Systems*, 30: 26